February 1, 1999

Genome-Wide Bioinformatic and Molecular Analysis of Introns in Saccharomyces Cerevisiae

Marc Spingola, *University of California, Santa Cruz*
Leslie Grate, *University of California, Santa Cruz*
David Haussler, *University of California, Santa Cruz*
Manuel Ares, *University of California, Santa Cruz*

RNA (1999), *5*:221–234. Cambridge University Press. Printed in the USA.

Genome-wide bioinformatic and molecular analysis of introns in *Saccharomyces cerevisiae*

MARC SPINGOLA,[1] LESLIE GRATE,[2] DAVID HAUSSLER,[2] and MANUEL ARES, JR.[1]

[1]Center for the Molecular Biology of RNA, Sinsheimer Laboratories, University of California–Santa Cruz, Santa Cruz, California 95064, USA
[2]Department of Computer and Information Sciences, Applied Sciences Building, University of California–Santa Cruz, Santa Cruz, California 95064, USA

ABSTRACT

Introns have typically been discovered in an *ad hoc* fashion: introns are found as a gene is characterized for other reasons. As complete eukaryotic genome sequences become available, better methods for predicting RNA processing signals in raw sequence will be necessary in order to discover genes and predict their expression. Here we present a catalog of 228 yeast introns, arrived at through a combination of bioinformatic and molecular analysis. Introns annotated in the *Saccharomyces* Genome Database (SGD) were evaluated, questionable introns were removed after failing a test for splicing in vivo, and known introns absent from the SGD annotation were added. A novel branchpoint sequence, AAUUAAC, was identified within an annotated intron that lacks a six-of-seven match to the highly conserved branchpoint consensus UACUAAC. Analysis of the database corroborates many conclusions about pre-mRNA substrate requirements for splicing derived from experimental studies, but indicates that splicing in yeast may not be as rigidly determined by splice-site conservation as had previously been thought. Using this database and a molecular technique that directly displays the lariat intron products of spliced transcripts (intron display), we suggest that the current set of 228 introns is still not complete, and that additional intron-containing genes remain to be discovered in yeast. The database can be accessed at http://www.cse.ucsc.edu/research/compbio/yeast_introns.html.

Keywords: branchpoint; hidden Markov model; intron database; intron display; splice site; splicing

INTRODUCTION

Much about eukaryotic genome organization has been revealed since the completion of the nuclear genome sequence for *Saccharomyces cerevisiae*. The set of genes carried by yeast numbers about 6,000 (Dujon, 1996; Goffeau et al., 1996; Mewes et al., 1997), and represents a model for understanding the evolution and sustenance of the eukaryotic cell. Unlike most eukaryotic genomes, the yeast genome has few introns. The introns are smaller on average, usually limited to one per gene in the few genes that carry them, are primarily located near the 5′ end of the gene, and have highly conserved splice sites and branchpoints (Fink, 1987; Woolford, 1989; Rymond & Rosbash, 1992). Why study yeast introns if their number and distribution are so skewed relative to what is observed in higher eukaryotes? Since spliceosomal introns are a uniquely eukaryotic genomic feature that obscures our ability to decipher the coding potential of genomes, it is useful to have a "simple" starting point for extension to more complex systems. Perhaps the few remaining introns in yeast hold the key to why eukaryotes have them. Despite the availability of the complete genome sequence, the annotation of yeast introns has been incomplete, and it has been difficult for biologists to access the data for analysis.

Since the last reviews of splicing in yeast (Woolford, 1989; Rymond & Rosbash, 1992), work by many laboratories has shown that an expanded set of signals is used to remove introns in yeast, thus making bioinformatic identification of all yeast introns more challenging. Yeast introns have been annotated using a program called EXPLORA (Kalogeropoulos, 1995), which relies heavily on open reading frame (ORF) information to predict introns. This feature makes it less likely to find many natural introns which have nonconsensus splice signals or do not lie within ORFs (Kalogeropoulos, 1995). Intron detection will be an even greater problem in the human genome, because of its higher intron content

Reprint requests to: Manuel Ares, Jr., Center for the Molecular Biology of RNA, Sinsheimer Laboratories, University of California–Santa Cruz, Santa Cruz, California 95064, USA; e-mail: ares@biology.ucsc.edu.

and more poorly conserved intron signals. In addition to computer analysis, molecular approaches will be necessary for confirming predictions and for finding new introns.

The first to attempt a global accounting of yeast introns were Rodriguez-Medina & Rymond (1994), who compared the in vitro translation products of RNA from a heat-treated, temperature-sensitive splicing mutant with an untreated control. After two-dimensional gel electrophoresis, they found that 2–3% of the many hundred protein spots detected required splicing for their translation. This number is consistent with the 228 annotated introns in 6,000 ORFs (~4%), but at the time this technique did not allow the identification of new introns.

We present here our initial efforts to identify and curate all yeast introns using a combination of bioinformatic and molecular approaches. We have used pattern searches for introns and developed a hidden Markov model (HMM; Krogh et al., 1994) to build an aligned set of introns and create a database of annotated introns that is easily searchable and available to the scientific community through the internet. This database has been purged of seven questionable annotated introns that failed a test for splicing. A number of introns not annotated in the *Saccharomyces* Genome Database (SGD) were added to our database. We have also developed a molecular method for detecting expressed yeast introns directly. This method, intron display, relies on a sequence of nucleotides present only in expressed introns. An oligonucleotide (a "branchmer") spanning the 2′-5′ junction of the lariat specifically primes reverse transcription of intron lariats that accumulate in a yeast mutant lacking the gene for the lariat debranching enzyme (Chapman & Boeke, 1991). When compared to our annotated set of introns, we find displayed introns whose lariat cDNA sizes do not correspond to any annotated introns, suggesting that the annotated set is incomplete and that more introns and genes remain hidden within the genome.

RESULTS AND DISCUSSION

Refining the database of annotated introns from SGD

We obtained the July 1997 and July 1998 releases of the annotated yeast genomic sequence from SGD. Since the database is annotated for exons rather than introns, we inspected all of the 228 ORFs annotated to have two or more exons. Fourteen of these ORFs had "introns" of only 2 nt between their exons, and all are −1 frameshift sites within Ty transposable elements (Wilson et al., 1986). These were removed. The nuclear gene for mitochondrial ribosomal protein L15 was annotated to have an intron of 153 nt because the sequence of the N-terminal protein fragment is downstream of the putative start codon (Grohmann et al., 1991). However this putative intron lacked an identifiable branchpoint or 3′ splice site. Further analysis indicated a sequencing error; the corrected sequence obviates the need to suppose an intron (Kitikawa et al., 1997). Based on these results, we eliminated YLR312w-a. Four of the remaining 213 ORFs have two introns for a total of 217 introns. Genes with two introns are *MATa1* (YCR097w), *YL8a/RPL7A* (YGL076c), *YL8b/RPL7B* (YPL198w), and *DYN2/SLC1* (YDR424c) (see, respectively, Miller, 1984; Mizuta et al., 1992; Mizuta et al., 1995; Dick et al., 1996).

We then added 17 confirmed introns that were not annotated in SGD. These include two introns in the U3A and U3B snoRNA genes (Myslinski et al., 1990), seven in the 5′ untranslated regions (5′UTRs) of mRNAs for *RPS8a* (YBL072c), *RPS8b* (YER102w), *RPS31* (YLR333c), *RPS31a* (YGR027c), *RPL30a* (YGL031c), *RPL30b* (YGR148c), *COX4* (YGL187c) (Logghe et al., 1994; Nieuwint et al., 1985; Mitra & Warner, 1984; Baronas-Lowell & Warner, 1990; Schneider & Guarente, 1987, respectively), and an additional six introns in *STO1* (YMR125w), *UBC8* (YEL012w), *UPF2* (YHR077c), *GCR1* (YPL075w), *MOB1* (YIL106w), and *ERV1* (YGR029w) (Uemura et al., 1996; Kaiser et al., 1994; Cui et al., 1995; Tornow & Santangelo, 1994; Luca & Winey, 1998; Lisowsky, 1996, respectively). Two new introns which have recently been identified reside in the genes for *YRA1* (YDR318w; Portman et al., 1997) and *REC114* (YMR133w; Malone et al., 1997). This brings the total to 234 introns in 230 genes. Molecular tests of questionable introns (see below) have lead us to remove six of these annotated introns for a total of 228. This refined intron database will be continuously updated and can be accessed at http://www.cse.ucsc.edu/research/compbio/yeast_introns.html.

Molecular tests challenge the authenticity of several unusual introns

We used the above set of intron sequences to create an initial HMM (Krogh et al., 1994; Hughey & Krogh, 1996; http://www.cse.ucsc.edu/research/compbio/sam.html), and aligned the set of annotated introns to the model. The initial HMM was trained without exon sequences, and identified the conserved 5′ splice site, the branchpoint, and the 3′ splice site. Inspection of the set of annotated introns aligned to the model revealed that several introns lacked obvious consensus-like sequences or had unusually small 5′ splice site-to-branchpoint distances. These include YDR535c and *CIN2* (YPL241c), which lack even a six-of-seven match to the branchpoint sequence UACUAAC, as well as five identical annotated introns found in the subtelomeric Y′ regions of various chromosomes (YBL111c, YML133c, YLL066c, YLL067c, and YHR218w), in which the distance between the 5′ splice site and branch point

is smaller than the suggested *S. cerevisiae* minimum of about 40 nt (Thompson-Jager & Domdey, 1987; Kohrer & Domdey, 1988).

We applied molecular tests to determine whether the above annotated introns were functional. First, we asked whether an intron could be detected by reverse transcription and polymerase chain reaction (RT-PCR) in RNA from vegetative haploid yeast cells. To exclude the possibility that the sequence containing the test intron was not transcribed under these conditions, we cloned a segment of the genome spanning the test intron into a high copy vector with a strong constitutive promoter (pGAC, Lesser & Guthrie, 1993a), and evaluated splicing of the RNA produced in haploid cells carrying the plasmids using direct primer extension (Fig. 1). YDR535c is annotated to have an intron of 255 nt to allow formation of a hypothetical ORF of 166 amino acids. This proposed intron lacks a good match to the branchpoint consensus sequence UACUAAC. Without the intron, YDR535c would code for a 49 amino acid protein. We were able to amplify the region by RT-PCR on total RNA, indicating that this portion of the genome is transcribed in haploid vegetative cells, yet we did not detect splicing of the annotated intron (data not shown). The region of the genome spanning the intron was cloned, expressed at high levels, and no splicing could be detected (Fig. 1, lane 4). Thus we conclude that YDR535c does not have this annotated intron and we removed it from our intron database.

Next we tested introns annotated in the Y′ elements, repeated sequences found in subtelomeric regions of yeast chromosomes. Some Y′ family members contain two large overlapping ORFs, one of which has homology to the RNA helicase family of proteins (Koonin, 1991). Louis & Haber (1992) analyzed the transcription and splicing of these elements using an RNase protection assay. Although they found that these loci are transcribed, they did not detect any splicing in the regions they tested. Five Y′ element family members are annotated to have nearly identical introns with an unusual but not unique GCAUGU 5′ splice site only 33 nt upstream of the proposed branchpoint. An RT-PCR assay verified their transcription in vegetative haploid cells, but no splicing of this intron family was observed (data not shown). We cloned a member of this family into the high copy expression plasmid, but could not detect any splicing of its transcript by primer extension (Fig. 1, lane 2). This observation and the unusually short 5′ splice site-to-branchpoint distance suggests that these five annotated introns are not real. We have removed them from our database.

In the case of the non-standard intron in the *CIN2* gene, splicing is observed. *CIN2* was isolated in screens for mutants hypersensitive to the antimicrotubule drug benomyl (Stearns et al., 1990), but there is no good match to the branchpoint consensus sequence in the annotated intron. We cloned a region spanning the annotated intron into pGAC, and splicing of the intron is detected (Fig. 1, lane 1). The abundance of *CIN2* transcripts produced from this plasmid allowed us to map the branchpoint directly. Primer extension using an oligonucleotide complementary to the second exon produces a strong stop at a cytidine residue within the intron (Fig. 2, lane 5). This stop is not observed if the RNA is treated with purified yeast debranching enzyme (Nam et al., 1994) prior to reverse transcription, and a new band corresponding to the 5′ end of the debranched lin-

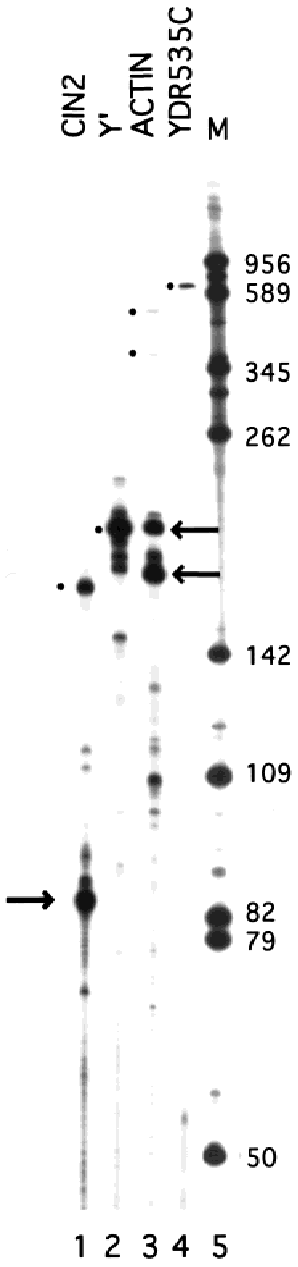

FIGURE 1. Splicing of yeast introns. Arrows indicate reverse transcription products from spliced RNA whereas dots indicate reverse transcription products corresponding to unspliced pre-mRNA. Lane 1: splicing of the *CIN2* intron; lane 2: Y′ element; lane 3: actin intron; lane 4: YDR535c; lane 5: size markers.

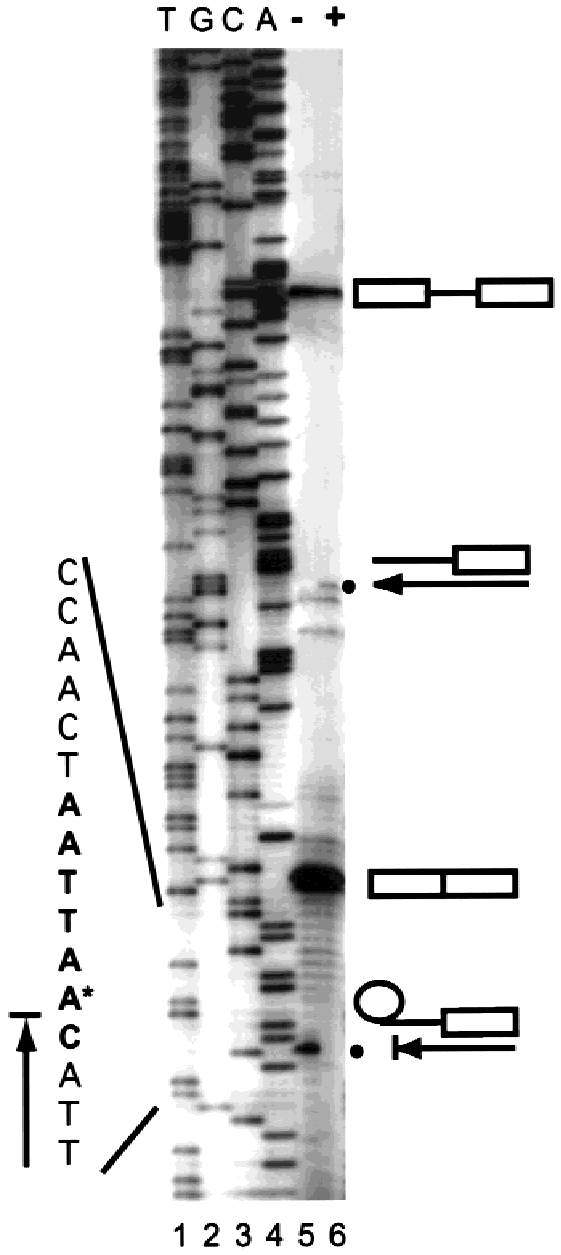

FIGURE 2. Branchpoint mapping of the *CIN2* intron by primer extension. Arrows indicate the bands corresponding to the reverse transcription products of lariat-2nd exon RNA and debranched lariat-2nd exon RNA. Reverse transcription products from spliced and unspliced RNA are also indicated on the right. The + refers to debranching at 30 °C and − to debranching at 4 °C. The sequence on the left indicates where the reverse-transcription stop occurs with the branchpoint sequence boldfaced. The branchpoint nucleotide is indicated with an asterisk. Lanes 1–4 are a set of dideoxynucleotide sequencing reactions; lane 5 is a debranching reaction incubated at 4 °C; lane 6 is a debranching reaction incubated at 30 °C.

ear intron appears (Fig. 2, lane 6). The nucleotide sequence upstream of the debranching-sensitive primer-extension stop is 5′-AAUUAA*C*-3′ (with the italicized C the point of reverse transcription stop; Fig. 2, lanes 1–4), indicating that the underlined A is the branched nucleotide (Ruskin & Green, 1985, 1990). The sequence AAUUAAC differs from the consensus UACUAAC at the first and third positions. These changes would disrupt pairing between the first position of UACUAAC and A39 of yeast U2, but allow a G-U pair between G37 of U2 and the third position of UACUAAC. This is the only example of a natural yeast intron with two deviations from the consensus branchpoint sequence.

Of the 228 introns currently in the database, approximately 60 lack experimental evidence for their existence, and only half of these 60 have homology to known proteins or protein families. This indicates that further analysis will be necessary to determine which of these represent true introns. The database can be searched remotely by gene name, sequence, 5′ splice site, branchpoint, or 3′ splice site. A table listing intron-containing genes is linked to the SGD, MIPS, and Proteome (YPD) web sites, and a page describing each intron with links to Medline is available.

The extent of natural splice site flexibility

Analysis of the 228 member intron set derived from the HMM alignment reveals that GUAUGU is the most commonly used 5′ splice site, with GUACGU second (Table 1). The remaining 5′ splice sites found so far are GUAUGA, GUAUGC, GCAUGU, GUAAGU, GUCAGU, and GUUCGU. Only the first and fifth positions (guanines) are invariant among all the annotated introns. Less variability is seen in the branchpoint sequences of the annotated set. Putative branchpoints were identified by the HMM, and in most cases remain to be confirmed. Of the 228 introns, 209 have UACUAAC branchpoints (Table 1). The remainder, with few exceptions, vary only at the first position of the branchpoint: GACUAAC, CACUAAC, and AACUAAC. There are three examples of a UGCUAAC branchpoint, two with UAUUAAC, and as shown above (Fig. 2), one with

TABLE 1. Types of various 5′ splice sites, branchpoints, and 3′ splice sites and counts of each used in *Saccharomyces cerevisiae*. Also listed are introns containing both an unusual 5′ splice site and an unusual branchpoint.

5′ splice site	number	branchpoint	number
GUAUGU	172	UACUAAC	209
GUACGU	26	CACUAAC	6
GUAUGA	11	GACUAAC	5
GUAUGC	8	UGCUAAC	3
GCAUGU	5	AACUAAC	2
GUAAGU	3	UAUUAAC	2
GUCAGU	2	AAUUAAC	1
GUUCGU	1		
combinations of unusual sites		3′ splice sites	
GUACGU+CACUAAC+YAG	2	UAG	116
GUACGU+GACUAAC+YAG	1	CAG	107
GUACGU+UACUAAC+AAG	1	AAG	5

AAUUAAC. The last four positions of the branchpoint sequence are invariant for all known examples. The vast majority of 3′ splice sites are YAG (Y = pyrimidine). The frequency of either U or C in the first position is nearly 50%. There are five examples of AAG as a 3′ splice site (Table 1).

As expected, the most common combinations of 5′ splice site with branchpoint are GUAUGU and UAC UAAC. However, combinations of less frequently used 5′ splice sites and branchpoints do occur (Table 1). For example, the gene *SUP46* uses a GUACGU 5′ splice site and a GACUAAC branchpoint (Vincent & Liebman, 1992). There are also two examples of GUACGU and CACUAAC combinations (in *RPS9A*/YPL081w and YBR219c), although we have noted that YBR219c also has a potential GACUAAC branchpoint. Of the five AAG 3′ splice sites, all use the most common branchpoint and 5′ splice site except one (YGL226c-a), which uses a GUACGU 5′ splice site (Reiss et al., 1997). The sample sizes of the rare splice site-containing introns are too small to determine whether the appearance of unusual combinations is significant.

The catalog of natural splice site and branchpoints identified in the culled intron database (Table 1) corroborates studies in which intron sequence elements were altered. For instance, mutations in the first and fifth positions of the 5′ splice site (invariant Gs for all naturally occurring yeast introns) adversely affect both the efficiency of splicing and specificity of the first catalytic step, whereas mutations at other positions have less severe effects (Jacquier et al., 1985; Parker & Guthrie, 1985; Fouser & Friesen, 1986; Seraphin et al., 1988; Seraphin & Rosbash, 1989; Kandels-Lewis & Seraphin, 1993; Lesser & Guthrie, 1993a). Branchpoint mutations at positions not found to vary naturally are the most detrimental to splicing efficiency (Jacquier et al., 1985; Jacquier & Rosbash, 1986; Rain & Legrain, 1997, Luukkonen & Seraphin, 1997). The weak effects of mutations at several positions in the 5′ splice site and branchpoints, and the existence of natural introns that use nonstandard signals (Table 1) indicate that the yeast splicing apparatus has flexibility in splice site usage. This flexibility is likely supported by additional nucleotides adjacent to the conserved sequences or pre-mRNA secondary structures (see below), and suggests that additional introns (and genes) may lie undiscovered or are incorrectly annotated in the yeast genome.

In general, the first YAG encountered after the branchpoint is expected to be the 3′ splice site (Fouser & Friesen, 1987; Smith et al., 1993; reviewed in Umen and Guthrie, 1995), however molecular experiments suggest an optimum distance (Patterson and Guthrie 1991, Frank & Guthrie, 1992; Brys & Schwer, 1996; Luukkonen & Seraphin, 1997; Zhang & Schwer, 1997). The annotated dataset of yeast introns shows a distribution of branchpoint to 3′ splice site distances with a mean of 39 nt (Fig. 3), in fair agreement with the experimental data of Luukkonen & Seraphin (1997), who found an optimum distance of 23 nt for a modified *RP51* intron. Many introns have a YAG sequence within 9 nt of the branchpoint and a distal YAG that represents the correct 3′ splice site, suggesting that YAG sequences too near the branchpoint are not used. Two uncharacterized introns are annotated to have their 3′ splice sites less than 9 nt from the branch, closer than the closest confirmed spacing (*MATa1* second intron is the shortest at 10 nt), but these may not be correctly annotated. Nine annotated introns have branchpoint to 3′ splice site distances in excess of 100 nt. Some of these have additional YAG sequences between the branchpoint and the annotated 3′ splice site. Studies using reporter constructs indicate that U-rich sequences (Patterson & Guthrie, 1991) or natural RNA structure downstream of the branchpoint (Deshler & Rossi, 1991) can influence 3′ splice site selection in yeast. Thus bioinformatic approaches to intron detection must consider multiple candidate 3′ splice sites and RNA structure.

Conserved context surrounding core splicing signals

We have used the refined database to generate additional hidden Markov models describing the known family of yeast introns (Fig. 4). Using 7 nt of exon at each end and the complete intron collection as a learning set, the model identifies the conserved 5′ splice site, branchpoint, and 3′ splice site sequences, illustrated

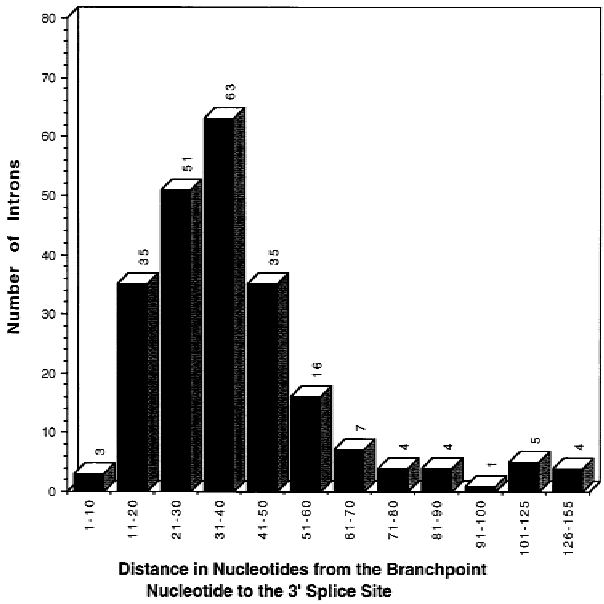

FIGURE 3. A histogram representing the distance in nucleotides from the branchpoint nucleotide to the 3′ splice site for all introns in the annotated set. Each column is grouped by 10 nt except the last two which are grouped by 25 nt.

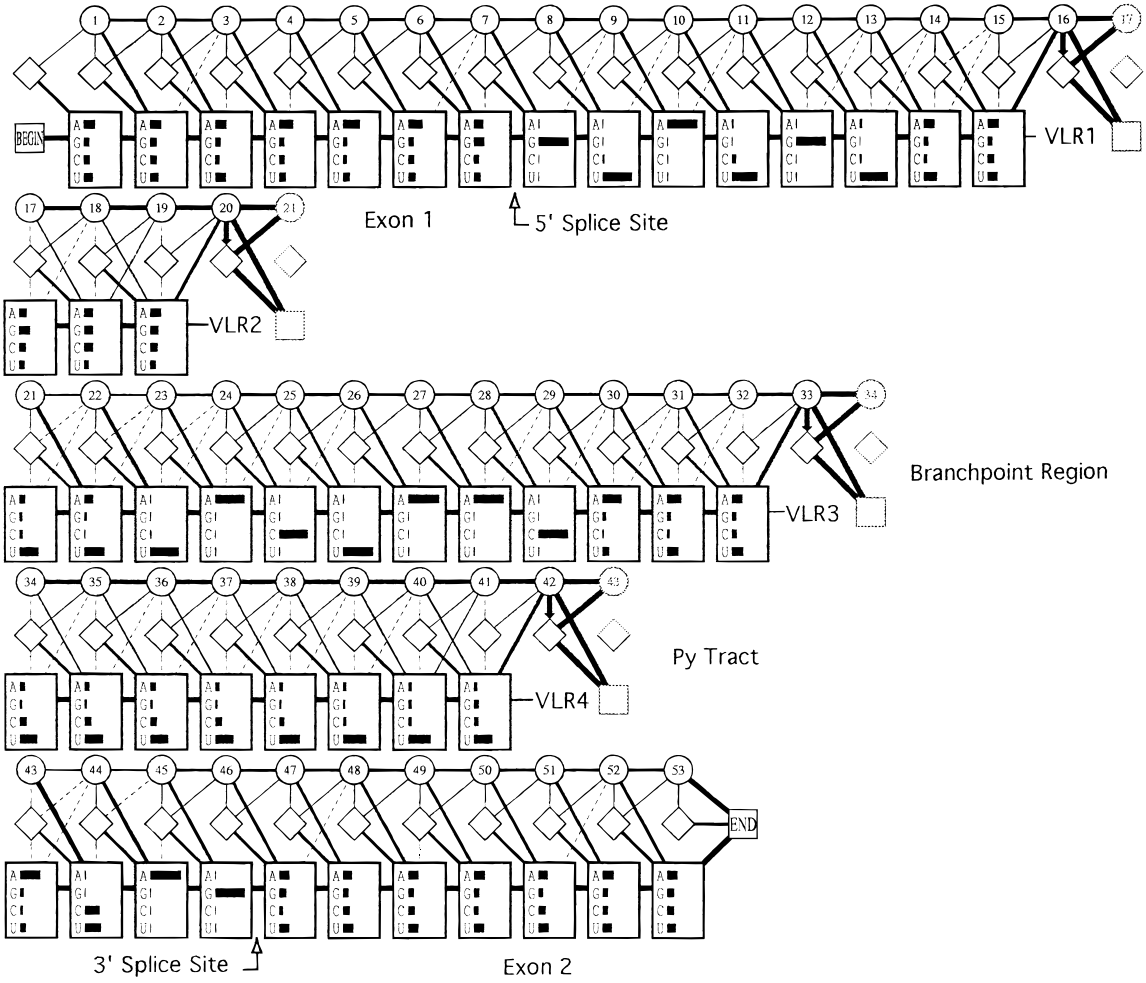

FIGURE 4. A hidden Markov model (HMM) for yeast introns. The annotated collection of introns was used as an initial learning set to generate this probabilistic model of yeast introns. A node in the model (node number indicated in the circle of each node) represents a nucleotide column in a multiple sequence alignment. As one proceeds through the model (starting at BEGIN) there is a defined probability of moving to the next defined state either within the same node or to the next node. The box represents the "consensus match" state, the diamond represents the "insert" state, and the circle and dashed lines represent the "delete" state. The thickness of the lines between states represent the probability of the indicated transition. Bars next to the nucleotides in the match state indicate the probability distribution of the nucleotides in that state. VLR1–4 indicate the positions of variable length regions (VLRs) which indicate positions at which arbitrary numbers of nucleotides may be inserted. Sequences can be compared to the model, and their scores can be evaluated relative to known introns.

by the bar graphs of nucleotide probabilities within the match states at each node (boxes in Fig. 4). Although not all yeast introns have one (Woolford, 1989; Rymond & Rosbash, 1992), a polypyrimidine tract is detected as a reasonably conserved sequence of U residues between the branchpoint and the 3′ splice site (Fig. 4). No other primary sequence motifs are globally conserved in the intron database.

Biased nucleotide distributions in the sequences adjacent to the conserved core signals are observed that may be of functional significance. In 41% of the annotated introns, a uridine follows the 5′ splice site (see Fig. 4). The significance of this observation is unclear since it does not extend the base pairing potential to either U1 or U6 snRNAs (but see below). A uridine also precedes the branchpoint sequence with a frequency of 65% and an A follows this sequence in nearly 66% of all 228 introns (Fig. 4), trends which have been identified previously (Parker et al., 1987; Rain & Legrain, 1997). The mechanistic roles of these nucleotides have not been tested rigorously; however the A at the end of the branchpoint could potentially base pair to U2 snRNA (Parker et al., 1987), and the preceding U may act by binding *trans*-acting factors, such as the Mud2 protein (Rain & Legrain, 1997; and reviewed in Staley & Guthrie, 1998) or the branchpoint binding protein BBP1 (Berglund et al., 1997). An adenosine precedes the YAG in 55% of the annotated introns.

The HMM also identifies a preference for AAAG preceding the 5′ splice site. This preference is not strong

and many introns lack these residues. On average, the three A nucleotides occur with a 50% frequency and the G with a 35% frequency. This observation is similar to that observed by Gilbert's group (Long et al., 1998). These nucleotides might be important for interacting with the U-rich loop of U5 snRNA (Seraphin & Rosbash, 1990; O'Keefe et al., 1996; and reviewed in Ares & Weiser, 1995) or enhancing binding of the first exon to PRP8 protein (Teigelkamp et al., 1995). In summary, the extended conservation surrounding these signals is represented by the model as AAAG/GUAUGUU for the 5′ splice site, UUUACUAACAA for the branchpoint, and AYAG for the 3′ splice site. The development of a probabilistic model of yeast introns will serve as a useful search tool as well as a reference point for the analysis of introns in other genomes. Using these data, mechanistic hypotheses (for example the possibility that A residues in the first exon may stabilize interactions with U5 or PRP8) can be developed and tested as well.

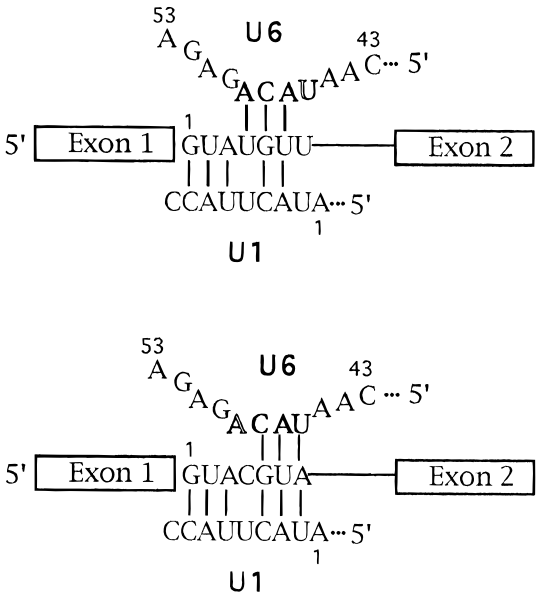

FIGURE 5. A model for the base pairing covariance between 5′ splice sites and U6 snRNA. In the upper illustration, ACA of U6 snRNA are base paired to the 5′ splice site and are boldfaced. In the bottom illustration CAU of U6 snRNA are base paired and boldfaced. The outlined nucleotides indicate nucleotides which can potentially base pair to one 5′ splice site but not the other.

Covariation between variant 5′ splice sites and a nearby residue

Additional information can be extracted by searching for significant covariations in the data. One example with a credible mechanistic explanation involves the known dual requirement of the 5′ splice site sequence for recognition by base pairing first by U1 (Seraphin et al., 1988; Siliciano & Guthrie, 1988) and then by U6 (Kandels-Lewis & Seraphin, 1993; Lesser & Guthrie, 1993b). The U1-5′ splice site pairing in most introns has a non-Watson–Crick U-U apposition at position 4 in the 5′ splice site. U6 pairs with conserved G5 as well as to positions 4 and 6 (Kandels-Lewis & Seraphin, 1993; Lesser & Guthrie, 1993b). In the variant 5′ splice site GUACGU, position 4 is changed from U to C, destabilizing the interaction with U6 (Fig. 5). Compensating for this loss is the appearance of an A residue 1 nt downstream of the 5′ splice site (position 7) in a significant proportion (73%) of introns lacking U at position 4 (Fig. 5). The covariation of 5′ splice site position 4 (U versus C) with position 7 (not A versus A) is very significant ($\chi^2 = 21$, degrees of freedom = 1, $P < 0.001$). The change from U to A at position 7 in this subset of introns probably stabilizes the U6 interaction and shifts its location by 1 nt. Other covariations that exist in the data may not be as easy to explain mechanistically, but their presence and description will allow better intron prediction from raw sequence data and may provide new insight into splicing mechanisms.

The distribution of yeast intron sizes

The length distribution of yeast introns, as seen in Figure 6, is bimodal with one peak around 100 nt and another near 400 nt. This trend has been observed since the earliest compilations of introns (Parker & Patterson, 1987; Hawkins, 1988; Woolford, 1989; Rymond & Rosbash, 1992; Rodriguez-Medina & Rymond, 1994). The largest intron is 1,001 nt (*DBP2*) and the smallest is 52 nt (*MATa1* intron 2). A major class of intron-containing genes is that encoding ribosomal proteins, translation factors, and factors involved in ribosome biogenesis. Approximately one third of all annotated introns reside in the genes coding for ribosomal proteins. These introns tend to be larger than introns residing in the genes of nonribosomal proteins. This phenomenon was also evident in earlier work (Parker & Patterson, 1987; Hawkins, 1988; Woolford, 1989; Rymond & Rosbash, 1992; Rodriguez-Medina & Rymond, 1994), and persists in the data derived from the complete genome. A few of the ribosomal introns might be large because they harbor snoRNAs. For example, *SNR39*, *SNR59*, *SNR44*, *SNR38*, U18, and U24 are encoded within the introns in the ribosomal protein or translation elongation factor genes *YL8A*, *YL8B*, *RPS22B*, *TEF4*, *TEF1*, and *BEL1/ASC1*, respectively (Bachellerie et al., 1995; Qu et al., 1995; Villa et al., 1998). Another possibility is that the presence of introns in these mRNAs may generally regulate their expression and subsequently affect ribosome biogenesis with an impact on growth rate (Li et al., 1996). Several examples of autoregulation through signals carried in introns are now known in yeast (Vilardell & Warner, 1997; Li et al., 1996; Barta & Iggo, 1995). Nonetheless, these two observations do not account for the abundance of introns in ribosomal protein genes or the unusual size distribution.

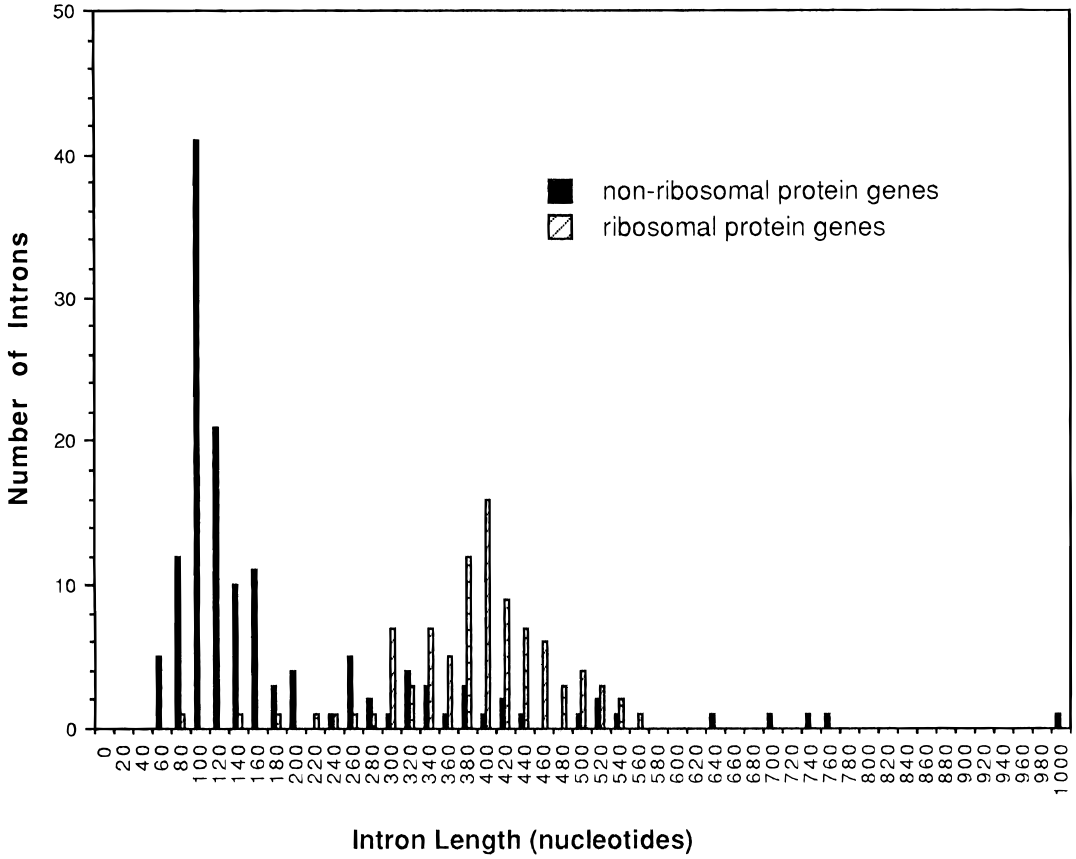

FIGURE 6. The length distribution of introns. Introns residing in genes for ribosomal proteins are indicated with hatched columns. All other introns are designated by black columns. Columns are grouped by 20 nt.

Larger introns, like those in ribosomal protein genes, have the capacity to form secondary structures which might bring the intron elements into closer proximity (Parker & Patterson, 1987). Disruption of these potential base pairings inhibits splicing efficiency (Newman 1989; Goguel & Rosbash, 1993; Libri et al., 1995; Charpentier & Rosbash, 1996; Howe & Ares, 1997), and can lead to exon skipping in multi-intron transcripts (Howe & Ares, 1997). Approximately 25% of the introns greater than 125 nt have obvious potential to form a helix that would bring the 5′ splice site close to the branchpoint (data not shown). Other introns (e.g., YDR450w, ribosomal protein S18A) that appear to lack a simple paired structure have the potential to form two helices, or more elaborate structures that have the same effect (data not shown). Thus no matter the gene in which they reside, the possibility that secondary structure within large introns may influence their splicing must be considered.

Molecular detection of expressed introns by intron display

It would be advantageous to the task of identifying the introns of a particular organism if they could be detected directly from extracts, regardless of their genomic organization. We have developed a simple primer-extension assay that can detect expressed intron lariats in yeast. The primer used for cDNA synthesis (a "branchmer") is typically short, about 15–20 nt, and is complementary to the sequence created when the 5′ splice site is joined to the branchpoint sequence of the lariat (spanning the 2′-5′ phosphodiester bond). Reverse transcription products from the branchmer are obtained only with RNA isolated from yeast lacking the gene for the debranching enzyme (*dbr1*Δ, Chapman & Boeke, 1991). This intron display technique is illustrated schematically in Figure 7A.

Reverse transcription of the actin lariat with the actin branchmer is shown in Figure 7B. RNA isolated from the *dbr1*Δ strain (Fig. 7B, lane 2), or the *dbr1*Δ strain carrying a high copy plasmid overexpressing a transcript containing the actin intron (Fig. 7B, lane 3), gives two major reverse-transcription products using the actin branchmer as a primer: a minor one of ~260 nt and a major one approximately 10 nt larger (Fig. 7B, lanes 2–3). The lengths of the products correspond to a complete cDNA of the lariat, and a complete lariat cDNA with 10 additional nucleotides expected if strand displacement synthesis were to occur past the 5′ end of the branchmer

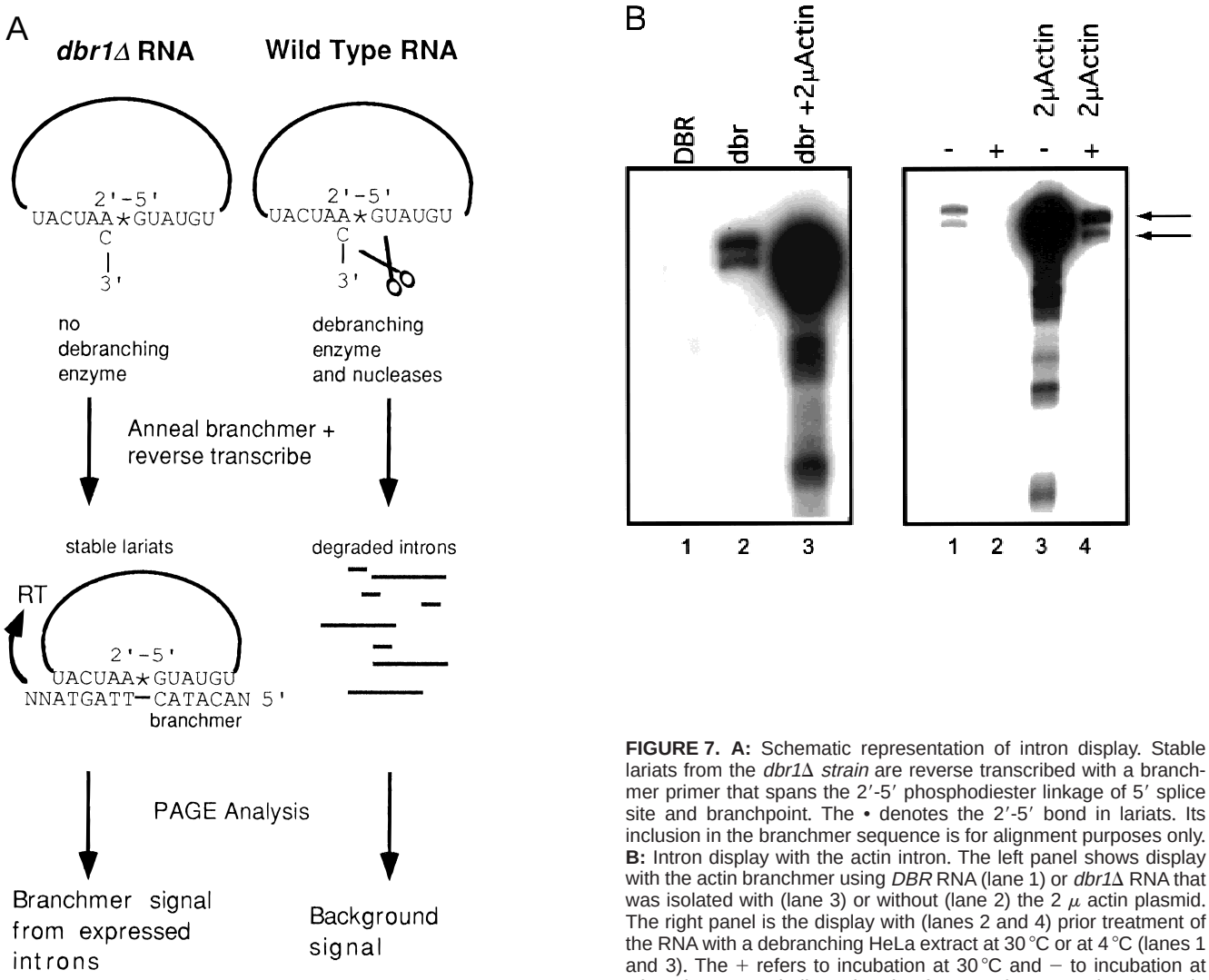

FIGURE 7. A: Schematic representation of intron display. Stable lariats from the *dbr1*Δ *strain* are reverse transcribed with a branchmer primer that spans the 2′-5′ phosphodiester linkage of 5′ splice site and branchpoint. The • denotes the 2′-5′ bond in lariats. Its inclusion in the branchmer sequence is for alignment purposes only. **B:** Intron display with the actin intron. The left panel shows display with the actin branchmer using *DBR* RNA (lane 1) or *dbr1*Δ RNA that was isolated with (lane 3) or without (lane 2) the 2 μ actin plasmid. The right panel is the display with (lanes 2 and 4) prior treatment of the RNA with a debranching HeLa extract at 30 °C or at 4 °C (lanes 1 and 3). The + refers to incubation at 30 °C and − to incubation at 4 °C. The arrows indicate bands of 260 and 270 nt. The 2 μ actin plasmid has a 6-nt deletion in its intron relative to the wild type.

up to the branchpoint 2′-5′ linkage. These products are not detectable when RNA from the wild-type *DBR1* strain is used as a template (Fig. 7B, lane 1). The template for these products is branched intron lariats rather than unspliced RNA based on the size of the cDNA and the observation that the signal decreases dramatically with prior treatment of the RNA with HeLa extract containing debranching activity (Ruskin & Green, 1985, 1990; Fig. 7B, right panel, lanes 1–4). Thus, expressed actin intron lariats can be detected directly in RNA from the *dbr1*Δ mutant strains, but not in wild-type *DBR1* strains using the actin branchmer.

Detection of nonoverlapping sets of introns with different branchmers

Although the 20-nt actin branchmer allows specific detection of a desired target lariat, genome-wide studies would benefit from detection of many expressed introns with as few branchmers as possible. We designed additional branchmers with the aim of detecting families of introns having the same 5′ splice site and branchpoint. Preliminary studies (not shown) indicate that branchmers with one randomized position on the 5′ end and two on the 3′ end (64-fold degenerate) allow efficient reverse transcription of lariats. We then tested several branchmers designed to display intron families (Fig. 8). The Type I branchmer (complementary to 5′NNUACUAA•GUAUGUN, N = any nucleotide) primes few products using wild-type *DBR1* RNA as a template (Fig. 8, lane 1), but primes many products using *dbr1*Δ RNA (Fig. 8, lane 2), suggesting that the branchmer detects many, if not all, of the expressed yeast introns spliced through the use of the standard 5′ splice site and branchpoint. The size distribution of cDNA products is bimodal, with the largest concentration around 300 nt and a second less abundant population around 100 nt (not shown), consistent with the anno-

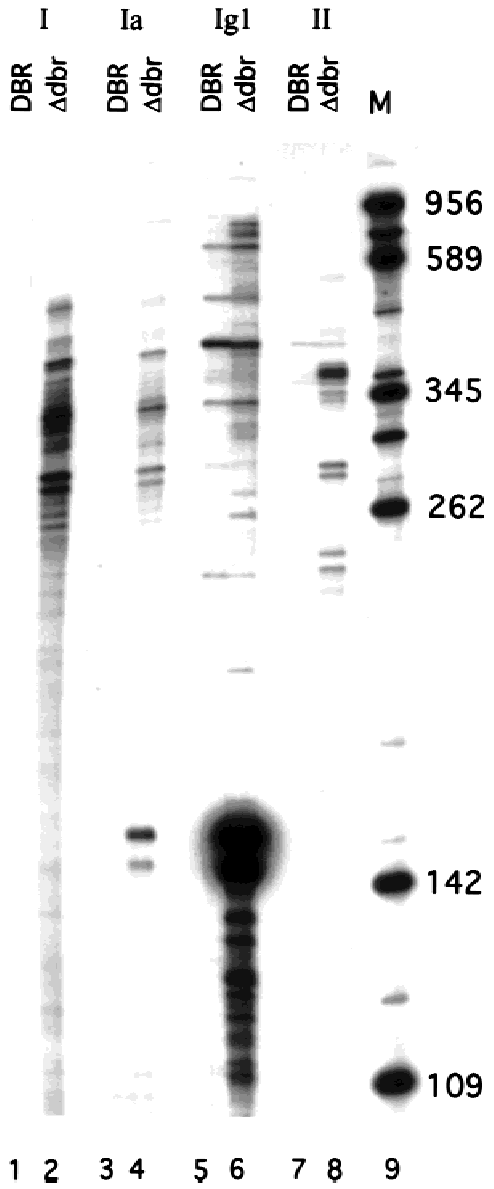

FIGURE 8. Intron display with branchmers designed to detect multiple expressed introns. Four branchmers were used separately for display with *dbr1*Δ or *DBR* RNA. Type I branchmer is complementary to 5′NNUACUAA•GUAUGUN (lane 1: *DBR* RNA; lane 2: *dbr1*Δ RNA). Type IA is complementary to 5′NNNACUAA•GUAUGUN (lane 3: *DBR* RNA; lane 4: *dbr1*Δ RNA). Type Ig1 is complementary to 5′NUGACUAA•GUAUGUN (lane 5: *DBR* RNA; lane 6: *dbr1*Δ RNA), and Type II is complementary to 5′NNUACUAA•GUACGUN (lane 7: *DBR* RNA; lane 8: *dbr1*Δ RNA). M (lane 9) contains size markers.

tated intron set (see Fig. 6). The Type II branchmer (complementary to 5′NNUACUAA•GUACGUN, the major variant 5′ splice site) produces a less complex, nonoverlapping set of products (Fig. 8, lanes 7 and 8), suggesting that the Type I and Type II branchmers each detect different subsets of expressed introns (Fig. 8, compare lane 2 with lane 8). The Type Ig1 branchmer is complementary to NUGACUAA•GUAUGUN (recognizes GACUAAC branchpoints preceded by a U, 16-fold degenerate) and produces many cDNAs unique to *dbr1*Δ RNA (Fig. 8, lanes 5 and 6). The set of bands from the Type Ig1 primer is clearly different from the Type I or II sets, indicating that different branchmers prime specifically on different sets of introns.

When we compare the introns displayed by the Type Ig1 branchmer, we find signals not accounted for by the annotated intron database, suggesting that a number of introns using the GACUAAC branchpoint remain undiscovered. There are five introns annotated with GAC UAAC branchpoints, and there are roughly 10 *dbr1*Δ-specific signals larger than the abundant products derived from the *SNR17A* intron at 145–150 nt (identity of the band confirmed by blotting nonradioactive branchmer cDNAs and radiolabeled U3 RNA; data not shown). One of the annotated five, *RPS9B*, has the GUACGU 5′ splice site and is not expected to be detected by the Type Ig1 primer. Two others, *SNR17B* and YBR230c, would give products shorter than 142 nt and are likely obscured by incomplete cDNAs primed on the abundant *SNR17A* lariats. Of the ~10 signals, only two can be accounted for, *YRA1* at about 746 nt and possibly YBR219c, whose branchpoint may have been incorrectly identified by the HMM (as CACUAAC), but which could use a GACUAAC branchpoint to produce a 388-nt lariat. Thus we can detect as many as eight undiscovered introns that utilize a GUAUGU 5′ splice site and a GACUAAC branchpoint. Since display depends on expression, this estimate is limited to genes expressed in vegetative *MATa* (haploid) cells.

Toward a complete accounting of yeast spliceosomal introns

The complete genome sequence of yeast has been available for more than two years, and as yet we have no comparable accounting of intron sequences. Although yeast has few introns, the genomes of other eukaryotes will present more serious interpretational challenges. By simple pattern searching or with more complex hidden Markov models, we have analyzed the annotated set of genes with more than two exons in the SGD. A number of annotations are unusual and molecular tests for splicing did not confirm their authenticity. We have concluded at this point that they are not bona fide introns. There still remain another 30 introns which lack any functional characterization, and another 30 which are annotated based solely on the resulting homology to known proteins or protein families once the predicted intron is removed from the hypothetical ORF. These observations stress the need to improve upon *in silico* methods of genome analysis, and more importantly, the requirement for molecular confirmation of computer-generated annotations. The improved annotation and accessibility of the yeast intron data should be a resource for future studies of splicing. The intron display technique outlined here should expedite mo-

lecular validation and genome-wide detection of expressed introns in yeast.

MATERIALS AND METHODS

Strains and oligonucleotides

A *dbr1*::*HIS3* strain of *S. cerevisiae* was created by transforming (Rothstein, 1991) with a *dbr1*::*HIS3* DNA fragment (a gift from J. Boeke). The resulting strain, HI201, is otherwise isogenic to the *DBR* strain HI227 (*MATa, GAL2, ura 3–52, leu 2–3,112, trp1, lys2*Δ*, his3*Δ*, pep4-3, prb1, prc1*). Gene disruption was verified by Southern analysis (data not shown). Oligonucleotides were synthesized by Cruachem and have the following sequences where N represents an equimolar mixture of all 4 nt: Actin branchmer: 5′TAGAACATACTTAG TACATG; Type I: 5′NACATACTTAGTANN; Type IA: 5′NA CATACTTAGTNNN; Type Ig1: 5′NACATACTTAGTCAN; and Type II: 5′NACGTACTTAGTANN. The *E. coli* strains DH5α and XL1-Blue were used for cloning.

Cloning

Genes with unusual or questionable introns from our annotated set were amplified by polymerase chain reaction (PCR; Saiki et al., 1988) from genomic preparations of yeast DNA (Philippsen et al., 1991). Primers were targeted to exon boundaries such that the entire intron and approximately 30 base pairs of each flanking exon were included (sequences available upon request). Amplifications were performed in 100 μL of 10 mM tris (pH 8.3 at 25 °C), 50 mM KCl, 1.5 mM $MgCl_2$, 0.01% gelatin with 0.5 μM of each primer, 500 ng DNA, and 1 unit of Vent DNA polymerase (New England Biolabs). Products were phenol/chloroform extracted, ethanol precipitated, and digested with restriction enzymes whose only sites were located in the 5′ end of the primers. The digested PCR products were then directionally cloned by standard methods (Sambrook et al., 1989) into a 2 μ pGAC (Lesser & Guthrie, 1993a) vector. The resulting plasmids used a glyceraldehyde 3-phosphate dehydrogenase promoter and a phosphoglucokinase terminator for transcription when introduced by transformation into *S. cerevisiae*. Transformants were plated on selective media, and RNA was isolated from liquid cultures by the hot phenol method as described in Zavanelli & Ares (1991).

Splicing analysis

Splicing of each of the cloned introns in vivo was analyzed by primer extension on total RNA using a 5′-radiolabeled primer (the PCR primer) complementary to the second exon followed by polyacrylamide gel electrophoresis. Typically, 5–10 μg of total RNA were annealed to 200 pg of 5′ ^{32}P-labeled primer in 10 μL of 50 mM Tris (pH 8.0)/80 mM KCl by first heating the sample to 90 °C for 2 min and then slowly cooling from 65 °C to 42 °C. To the annealed mix were added 10 μL of 50 mM Tris (pH 8.0), 80 mM KCl , 20 mM $MgCl_2$, 10 mM dithiothreitol (DTT), 0.2 mM dNTP, 100 μg actinomycin D, and 20 U of AMV reverse transcriptase. Reactions were incubated at 42 °C for 30–60 min and terminated by adding 10 μL of 0.6 M NaOAc (pH 5.0), 60 mM EDTA, 1 μg/mL RNase A with further incubation for 10 min at 42 °C. To this were added 10 μL of 0.6 M NaOAc (pH 5.0), 0.2% SDS, 20 μg proteinase K and it was incubated for another 10 min at 42 °C. Samples were then precipitated on dry ice after the addition of 3 volumes of ethanol. Precipitates were centrifuged 5 min at 14,000 rpm, the supernatants were removed, and the pellets were dried. Pellets were redissolved in 3 μL of water and 3 μL of formamide/20 mM EDTA/0.25% bromophenol blue and xylene cyanol and then denatured for 5 min at 90 °C. The reverse transcription products were electrophoresed through 0.4-mm thick, 20-cm long 6% polyacrylamide, 7.5 M urea gels. Gels were dried and exposed to X-ray film overnight at −70 °C.

Potential splicing of the YDR535c and the Y′ introns was also assessed by reverse transcription PCR (RT-PCR) using RNA from the *DBR1* strain. All steps were performed as above, except that the RNA was treated with DNAse I prior to reverse transcription, and half an RT reaction was directly used for PCR after inactivating the reverse transcriptase at 65 °C for 10 min. The RNase, protease, and precipitation steps were omitted. The PCR products were analyzed by conventional agarose electrophoresis.

Branchpoint mapping

The branchpoint of the *CIN2* intron was mapped by primer extension using total RNA isolated from *dbr1*Δ containing pGCIN2. The reverse transcription products were electrophoresed parallel to a dideoxynucleotide sequencing reaction (Sanger et al., 1977) of the plasmid pGCIN2 using the same phosphorylated primer (second exon) as for reverse transcription. The 2′–5′ linkage at the branchpoint nucleotide results in a primer-extension stop at the nucleotide preceding the branchpoint nucleotide. RNA was also treated with purified recombinant debranching enzyme (a gift from the Boeke laboratory) to debranch the *CIN2* lariat. Ten micrograms of RNA were added to 0.5 μL of enzyme (93 mg/mL) in 15 μL of 20 mM HEPES-KOH (pH 8.0), 125 mM KCl, 0.5 mM $MgCl_2$, 1 mM DTT, 15 mM EDTA, and 10% glycerol and incubated at 30 °C for 30 min. Samples were then phenol/chloroform extracted and ethanol precipitated in 0.3 M NaOAc (pH 5.0) and 3 volumes of ethanol. The dried pellet was used for primer extension as above.

HeLa extracts were also used to debranch total RNA by the method of Ruskin & Green (1985, 1990).

Intron display

With the exception of pilot experiments with the actin intron, total RNA was isolated from yeast lacking any plasmids. RNA from *DBR1* and *dbr1*Δ strains of *S. cerevisiae* was reverse transcribed as above using a 5′ radiolabeled branchmer primer. All products were electrophoresed as above.

Bioinformatics

The July 1997 and July 1998 data from the SGD were obtained at http://genome-www.stanford.edu/Saccharomyces. All ORFs containing two or more exons were extracted and inspected for correctness. The corrected set of 228 introns was analyzed with a hidden Markov model (using SAM, available at http://www.cse.ucsc.edu/research/compbio/sam.html)

to extract the globally conserved primary sequence motifs and to create an alignment of the introns.

NOTE ADDED IN PROOF

A novel yeast 5′ splice site lacking a G at position 5 was recently reported by Leu et al. (Leu JY, Chua PR, Roeder GS. 1998. The meiosis-specific Hop2 protein of *S. cerevisiae* ensures synapsis between homologous chromosomes. *Cell 94*:375–386.)

ACKNOWLEDGMENTS

We thank Chris Greer, Steve Mount, and Sandy Wells for previewing the web site, and Jef Boeke for clones and purified DBR enzyme. We also thank Carrie Davis for help with the web site and Michelle Haynes Pauling, Ken Howe, and Bill Grundy for critical comments on the manuscript. This work was supported by a grant from the National Cancer Institute: R21 CA77813.

Received September 14, 1998; returned for revision October 9, 1998; revised manuscript received October 20, 1998

REFERENCES

Ares M Jr, Weiser B. 1995. Rearrangement of snRNA structure during assembly and function of the spliceosome. *Prog Nucleic Acid Res Mol Bio 50*:131–159.

Bachellerie JP, Michot B, Nicoloso M, Balakin A, Ni J, Fournier MJ. 1995. Antisense snoRNAs: A family of nucleolar RNAs with long complementarities to rRNA. *Trends Biochem Sci 20*:261–264.

Baronas-Lowell DM, Warner JR. 1990. Ribosomal protein L30 is dispensable in the yeast *Saccharomyces cerevisiae. Mol Cell Biol 10*:5235–5243.

Barta I, Iggo R. 1995. Autoregulation of expression of the yeast Dbp2p 'DEAD-box' protein is mediated by sequences in the conserved DBP2 intron. *EMBO J 14*:3800–3808.

Berglund JA, Chua K, Abovich N, Reed R, Rosbash M. 1997. The splicing factor BBP interacts specifically with the pre-mRNA branchpoint sequence UACUAAC. *Cell 89*:781–787.

Brys A, Schwer B. 1996. Requirement for SLU7 in yeast pre-mRNA splicing is dictated by the distance between the branchpoint and the 3′ splice site. *RNA 2*:707–717.

Chapman KB, Boeke JD. 1991. Isolation and characterization of the gene encoding yeast debranching enzyme. *Cell 65*:483–492.

Charpentier B, Rosbash M. 1996. Intramolecular structure in yeast introns aids the early steps of in vitro spliceosome assembly. *RNA 2*:509–522.

Cui Y, Hagan KW, Zhang S, Peltz SW. 1995. Identification and characterization of genes that are required for the accelerated degradation of mRNAs containing a premature translational termination codon. *Genes & Dev 9*:423–436.

Deshler JO, Rossi JJ. 1991. Unexpected point mutations activate cryptic 3′ splice sites by perturbing a natural secondary structure within a yeast intron. *Genes & Dev 1*:445–455.

Dick T, Surana U, Chia W. 1996. Molecular and genetic characterization of SLC1, a putative *Saccharomyces cerevisiae* homolog of the mammalian cytoplasmic dynein light chain 1. *Mol Gen Genet 251*:38–43.

Dujon B. 1996. The yeast genome project: What did we learn? *Trends Genet 12*:263–270.

Fink GR. 1987. Pseudogenes in yeast. *Cell 42*:355–367.

Fouser LA, Friesen JD. 1986. Mutations in a yeast intron demonstrate the importance of specific conserved nucleotides for the two stages of nuclear mRNA splicing. *Cell 45*:81–93.

Fouser LA, Friesen JD. 1987. Effects on mRNA splicing of mutations in the 3′ region of the *Saccharomyces cerevisiae* actin intron. *Mol Cell Biol 7*:225–230.

Frank D, Guthrie C. 1992. An essential splicing factor, SLU7, mediates 3′ splice site choice in yeast. *Genes & Dev 6*:2112–2124.

Goffeau A, Barrel BG, Bussey H, Davis RW, Dujon B, Feldmann H, Galibert F, Hoheisel JD, Jacq C, Johnston M, Louis EJ, Mewes HW, Murakami Y, Philippsen P, Tettelin H, Oliver SG. 1996. Life with 6000 genes. *Science 274*:546–567.

Goguel V, Rosbash M. 1993. Splice site choice and splicing efficiency are positively influenced by pre-mRNA intramolecular base pairing in yeast. *Cell 72*:893–901.

Grohmann L, Graack H, Kruft V, Choli T, Goldschmidt-Reisin S, Kitakawa M. 1991. Extended N-terminal sequencing of proteins of the large ribosomal subunit from yeast mitochondria. *FEBS Lett 284*:51–56.

Hawkins JD. 1988. A survey on intron and exon lengths. *Nucleic Acids Res 16*:9893–9908.

Howe KJ, Ares M Jr. 1997. Intron self-complementarity enforces exon inclusion in a yeast pre-mRNA. *Proc Natl Acad Sci USA 94*:12467–12472.

Hughey R, Krogh A. 1996. Hidden Markov models for sequence analysis: Extension and analysis of the basic method. *Cabios 12*:95–107.

Jacquier A, Rodriguez JR, Rosbash M. 1985. A quantitative analysis of the effects of 5′ junction and TACTAAC box mutants and mutant combinations on yeast mRNA splicing. *Cell 43*: 423–430.

Jacquier A, Rosbash M. 1986. RNA splicing and intron turnover are greatly diminished by a mutant yeast branchpoint. *Proc Natl Acad Sci USA 83*:5835–5839.

Kaiser P, Suefert W, Hofferer L, Kofler B, Sachsenmaier C, Herzog H, Jentsch S, Schweiger M, Schneider R. 1994. A human ubiquitin-conjugating enzyme homologous to yeast UBC8. *J Biol Chem 269*:8797–8802.

Kalogeropoulos A. 1995. Automatic intron detection in nuclear DNA sequences of *Saccharomyces cerevisiae. Yeast 11*:555–565.

Kandels-Lewis S, Seraphin B. 1993. Role of U6 snRNA in 5′ splice site selection. *Science 262*:2035–2039.

Kitakawa M, Graack H, Grohmann L, Goldschmidt-Reisin S, Herfurth E, Wittmann-Liebold B, Nishimura T, Isono K. 1997. Identification and characterization of the genes for mitochondrial ribosomal protiens of *Saccharomyces cerevisiae. Eur J Biochem 245*:449–456.

Kohrer K, Domdey H. 1988. Splicing and spliceosome formation of the yeast MATa1 transcript require a minimum distance from the 5′ splice site to the internal branch acceptor site. *Nucleic Acids Res 16*:9457–9475.

Koonin EV. 1991. Similarities in RNA helicases. *Nature 352*:290–293.

Krogh A, Brown M, Mian IS, Sjolander K, Haussler D. 1994. Hidden Markov models in computational biology. Applications to protein modeling. *J Mol Biol 235*:1501–1531.

Lesser CF, Guthrie C. 1993a. Mutational analysis of pre-mRNA splicing in *Saccharomyces cerevisiae* using a sensitive new reporter gene, *CUP1. Genetics 133*:851–863.

Lesser CF, Guthrie C. 1993b. Mutations in U6 snRNA that alter splice site specificity: Implications for the active site. *Science 262*:1982–1988.

Li B, Vilardell J, Warner JR. 1996. An RNA structure involved in the feedback regulation of splicing and of translation is critical for biological fitness. *Proc Natl Acad Sci USA 93*:1596–1600.

Libri D, Stutz F, McCarthy T, Rosbash M. 1995. RNA structural patterns and splicing: Molecular basis for an RNA-based enhancer. *RNA 1*:425–436.

Lisowsky T. 1996. Removal of an intron with a unique 3′ branch site creates an amino-terminal protein sequence directing the scERV1 gene product to mitochondria. *Yeast 12*:1501–1510.

Logghe M, Molemans F, Fiers W, Contreras R. 1994. The two genes encoding yeast ribosomal protein S8 reside on different chromo-

somes, and are closely linked to the hsp70 stress protein genes SSA3 and SSA4. *Yeast 10*:1093–1100.
Long M, de Souza SJ, Gilbert W. 1998. The yeast splice site revisited: New exon consensus from genomic analysis. *Cell 91*:739–740.
Louis EJ, Haber JE. 1992. The structure and function of subtelomeric Y′ repeats in *Saccharomyces cerevisiae. Genetics 133*:559–574.
Luca FC, Winey M. 1998. MOB1, an essential yeast gene required for completion of mitosis and maintenance of ploidy. *Mol Biol Cell 9*:29–46.
Luukkonen BGM, Seraphin B. 1997. The role of branchpoint-3′ splice site spacing and interaction between intron terminal nucleotides in 3′ splice site selection in *Saccharomyces cerevisiae. EMBO J 16*:779–792.
Malone RE, Pittman DL, Nau JJ. 1997. Examination of the intron in the meiosis-specific recombination gene REC114 in *Saccharomyces. Mol Gen Genet 255*:410–419.
Mewes HW, Albermann K, Bahr M, Frishman D, Gleissner A, Hani J, Heumann K, Kleine K, Maierl A, Oliver SG, Pfeiffer F, Zollner A. 1997. Overview of the yeast genome. *Nature 387*:7–8.
Miller AM. 1984. The yeast MATa1 gene contains two introns. *EMBO J 3*:1061–1065.
Mitra G, Warner JR. 1984. A yeast ribosomal protein gene whose intron is in the 5′ leader. *J Biol Chem 259*:9218–9224.
Mizuta K, Hashimoto T, Otaka E. 1992. Yeast ribosomal proteins: XIII. *Saccharomyces cerevisiae YL8A* gene, interrupted with two introns, encodes a homolog of mammalian L7. *Nucleic Acids Res 20*:1011–1016.
Mizuta K, Hashimoto T, Otaka E. 1995. The evolutionary relationships between homologs of ribosomal YL8 protein and YL8-like proteins. *Curr Genet 28*:19–25.
Myslinski E, Segault V, Branlant C. 1990. An intron in the genes for U3 small nucleolar RNAs of the yeast *Saccharomyces cerevisiae. Science 247*:1213–1216.
Nam K, Hudson RHE, Chapman K, Ganeshan K, Damha M, Boeke JD. 1994. Yeast lariat debranching enzyme. Substrate and sequence specificity. *J Biol Chem 269*:20613–20621.
Newman A. 1989. Specific accessory sequences in *Saccharomyces cerevisiae* introns control the assembly of pre-mRNAs into spliceosomes. *EMBO J 6*:3833–3839.
Nieuwint RT, Molenaar CM, van Bommel JH, van Raamsdonk-Duin MM, Mager WH, Planta RJ. 1985. The gene for yeast ribosomal protein S31 contains an intron in the leader sequence. *Curr Genet 10*:1–5.
O'Keefe RT, Norman C, Newman AJ. 1996. The invariant U5 snRNA loop 1 sequence is dispensable for the first catalytic step of pre-mRNA splicing in yeast. *Cell 86*:679–689.
Parker R, Guthrie C. 1985. A point mutation in the conserved hexanucleotide at a yeast 5′ splice junction uncouples recognition, cleavage, and ligation. *Cell 41*:107–118.
Parker R, Patterson B. 1987. Architecture of fungal introns: Implications for spliceosome assembly. In: Dudock B, Inouye M, eds. *Molecular biology of RNA, new perspectives*. New York: Academic Press. pp 133–149.
Parker R, Siliciano PG, Guthrie C. 1987. Recognition of the TAC TAAC box during mRNA splicing in yeast involves base pairing to the U2-like snRNA. *Cell 41*:107–118.
Patterson B, Guthrie C. 1991. A U-rich tract enhances usage of an alternative 3′ splice site in yeast. *Cell 64*:181–187.
Philippsen P, Stotz A, Scherf C. 1991. DNA of *Saccharomyces cerevisiae. Methods Enzymol 194*:168–182.
Portman DS, O'Connor JP, Dreyfuss G. 1997. YRA1, an essential *Saccharomyces cerevisiae* gene, encodes a novel nuclear protein with RNA annealing activity. *RNA 3*:527–537.
Qu LH, Henry Y, Nicoloso M, Michot B, Azum MC, Renalier MH, Caizergues-Ferrer M, Bachellerie JP. 1995. U24, a novel intron-encoded small nucleolar RNA with two 12 nt long, phylogenetically conserved complementarities to 28S rRNA. *Nucleic Acids Res 23*:2669–2676.
Rain J-C, Legrain P. 1997. In vivo commitment to splicing in yeast involves the nucleotide upstream from the branch site conserved sequence and the Mud2 protein. *EMBO J 16*:1759–1771.
Reiss G, te Heesen S, Gilmore R, Zufferey R, Aebi M. 1997. A specific screen for oligosaccharyltransferase mutations identifies the 9 kDa OST5 protein required for optimal activity in vivo and in vitro. *EMBO J 16*:1164–1172.
Rodriguez-Medina JR, Rymond BC. 1994. Prevalence and distribution of introns in non-ribosomal protein genes of yeast. *Mol Gen Genet 243*:532–539.
Rothstein R. 1991. Targeting, disruption, replacement, and allele rescue: Integrative DNA transformation in yeast. *Methods Enzymol 194*:281–301.
Ruskin B, Green MR. 1985. An RNA processing activity that debranches lariats. *Science 229*:135–140.
Ruskin B, Green MR. 1990. RNA lariat debranching enzyme as a tool for analyzing RNA structure. *Methods Enzymol 181*:180–188.
Rymond B, Rosbash M. 1992. Yeast pre-mRNA splicing. In: Broach JR, Pringle J, Jones EW, eds. *The molecular and cellular biology of the yeast Saccharomyces cerevisiae,* Vol. 2. Cold Spring Harbor, New York: Cold Spring Harbor Laboratory Press. pp 143–192.
Saiki RK, Gelfand D, Stoffel S, Scharf SJ, Higuchi R, Horn GT, Mullis KB, Erlich HA. 1988. Primer-directed enzymatic amplification of DNA with a thermostable DNA polymerase. *Science 239*:487–491.
Sambrook J, Fritsch EF, Maniatis T. 1989. *Molecular cloning: A laboratory manual*. 2nd ed. Cold Spring Harbor, New York: Cold Spring Harbor Laboratory Press.
Sanger F, Nicklen S, Coulson AR. 1977. DNA sequencing with chain terminating inhibitors. *Proc Natl Acad Sci USA 74*:5463–5469.
Schneider JC, Guarente L. 1987. The untranslated leader of nuclear COX4 gene of *Saccharomyces cerevisiae* contains an intron. *Nucleic Acids Res 15*:3515–3529.
Seraphin B, Kretzner L, Rosbash M. 1988. A U1 snRNA:pre-mRNA base pairing interaction is required early in yeast spliceosome assembly but does not uniquely define the 5′ cleavage site. *EMBO J 7*:2533–2538.
Seraphin B, Rosbash M. 1989. Mutational analysis of the interactions between U1 small nuclear RNA and pre-mRNA of yeast. *Gene 82*:145–151.
Seraphin B, Rosbash M. 1990. Exon mutations uncouple 5′ splice site selection from U1 snRNA pairing. *Cell 63*:619–629.
Siliciano PG, Guthrie C. 1988. 5′ splice site selection in yeast: Genetic alterations in base-pairing with U1 reveal additional requirements. *Genes & Dev 2*:1258–1267.
Smith CW, Chu TT, Nadal-Ginard B. 1993. Scanning and competition between AGs are involved in 3′ splice site selection in mammalian introns. *Mol Cell Biol 13*:4939–4952.
Staley JP, Guthrie C. 1998. Mechanical devices of the spliceosome: Motors, clocks, springs, and things. *Cell 92*:315–326.
Stearns T, Hoyt MA, Botstein D. 1990. Yeast mutants sensitive to antimicrotubule drugs define three genes that affect microtubule function. *Genetics 124*:251–262.
Teigelkamp S, Newman AJ, Beggs JD. 1995. Extensive interactions of PRP8 protein with the 5′ and 3′ splice sites during splicing suggests a role in stabilization of exon alignment by U5 snRNA. *EMBO J 14*:2602–2612.
Thompson-Jager S, Domdey H. 1987. Yeast pre-mRNA splicing requires a minimum distance between the 5′ splice site and the internal branch acceptor site. *Mol Cell Biol 7*:4010–4016.
Tornow J, Santangelo G. 1994. The GCR1 gene of *Saccharomyces cerevisiae* is a split gene with an unusually long intron. *Genetics 138*:493–505.
Uemura H, Pandit S, Jigami Y, Sternglanz R. 1996. Mutations in GCR3, a gene involved in the expression of glycolytic genes in *Saccharomyces cerevisiae*, suppress the temperature-sensitive growth of hpr1 mutants. *Genetics 142*:1095–1013.
Umen JG, Guthrie C. 1995. The second catalytic step of pre-mRNA splicing. *RNA 1*:869–885.
Vilardell J, Warner J. 1997. Ribosomal protein L32 of *Saccharomyces cerevisiae* influences both the splicing of its own transcript and the processing of rRNA. *Mol Cell Biol 17*:1959–1965.
Villa T, Ceradini F, Presutti C, Bozzoni I. 1998. Processing of the intron-encoded U18 small nucleolar RNA in the yeast *Saccharomyces cerevisiae* relies on both exo- and endonucleolytic activities. *Mol Cell Biol 18*:3376–3383.

Vincent A, Liebman SW. 1992. The yeast omnipotent suppressor SUP46 encodes a ribosomal protein which is a functional and structural homolog of the *Escherichia coli* S4 ram protein. *Genetics 132*:375–386.

Wilson W, Malim MH, Mellor J, Kingsman AJ, Kingsman SM. 1986. Expression strategies of the yeast retrotransposon Ty: A short sequence directs ribosomal frameshifting. *Nucleic Acids Res 14*:7001–7016

Woolford JL Jr. 1989. Nuclear pre-mRNA splicing in yeast. *Yeast 5*:439–457.

Zavanelli MI, Ares M Jr. 1991. Efficient association of U2 snRNPs with pre-mRNA requires an essential U2 RNA structural element. *Genes & Dev 5*:2521–2533.

Zhang X, Schwer B. 1997. Functional and physical interaction between the yeast splicing factors Slu7 and Prp18. *Nucleic Acids Res 25*:2146–2152.