Nucleic Acids Research, 2003, Vol. 31, No. 4 ***1121–1135***
DOI: 10.1093/nar/gkg213

Molecular evolution of eukaryotic genomes: hemiascomycetous yeast spliceosomal introns

Elisabeth Bon, Serge Casaregola, Gaëlle Blandin[1], Bertrand Llorente[1], Cécile Neuvéglise, Martin Munsterkotter[2], Ulrich Guldener[2], Hans-Werner Mewes[2], Jacques Van Helden[3], Bernard Dujon[1] and Claude Gaillardin*

Laboratoire de Génétique Moléculaire et Cellulaire CNRS-INRA, Institut National Agronomique Paris-Grignon, F-78850 Thiverval-Grignon, France, [1]Unité de Génétique Moléculaire des Levures, Institut Pasteur/URA 2171 CNRS et UFR 927, Université Pierre et Marie Curie, Institut Pasteur, 25 rue du Docteur Roux, F-75724 Paris Cedex, France, [2]MIPS, Institut für Bioinformatik, GSF-Forschungszentrum für Umwelt und Gesundheit, GmbH-, Ingolstädter Landstraβe 1, D-85764 Neuherberg, Germany and [3]Unité de Conformation des Macromolécules Biologiques, Université Libre de Bruxelles, CP 263, Boulevard du Triomphe, B-1050 Bruxelles, Belgium

Received December 11, 2002; Accepted December 19, 2002

ABSTRACT

As part of the exploratory sequencing program *Génolevures*, visual scrutinisation and bioinformatic tools were used to detect spliceosomal introns in seven hemiascomycetous yeast species. A total of 153 putative novel introns were identified. Introns are rare in yeast nuclear genes (<5% have an intron), mainly located at the 5′ end of ORFs, and not highly conserved in sequence. They all share a clear non-random vocabulary: conserved splice sites and conserved nucleotide contexts around splice sites. Homologues of metazoan snRNAs and putative homologues of SR splicing factors were identified, confirming that the spliceosomal machinery is highly conserved in eukaryotes. Several introns' features were tested as possible markers for phylogenetic analysis. We found that intron sizes vary widely within each genome, and according to the phylogenetic position of the yeast species. The evolutionary origin of spliceosomal introns was examined by analysing the degree of conservation of intron positions in homologous yeast genes. Most introns appeared to exist in the last common ancestor of present day yeast species, and then to have been differentially lost during speciation. However, in some cases, it is difficult to exclude a possible sliding event affecting a pre-existing intron or a gain of a novel intron. Taken together, our results indicate that the origin of spliceosomal introns is complex within a given genome, and that present day introns may have resulted from a dynamic flux between intron conservation, intron loss and intron gain during the evolution of hemiascomycetous yeasts.

INTRODUCTION

The removal of introns from pre-mRNAs involves similar molecular mechanisms in all eukaryotes (1–3). Most of the time, this process is mediated by the 'U2-dependent spliceosome', a ribonucleoprotein complex composed of five phylogenetically conserved small nuclear RNAs (snRNAs) (U1, U2, U4, U5 and U6) assembled into small ribonucleoprotein particles (snRNPs), plus a set of transiently associated proteins (RNA helicases, SR splicing factors, etc.). This large complex is a highly dynamic structure initially thought to be assembled onto pre-mRNA introns in a stepwise manner (4). Recent work, however, has brought new insights into the spliceosome assembly pathway in eukaryotes (5,6). It notably demonstrated that yeast spliceosomal snRNPs pre-assemble into a 45S penta-snRNP complex prior to binding the pre-mRNA substrate (6). The spliceosome is organised, at least in part, by complementary RNA base pairing between the snRNAs and some specific target sites located at the 5′ and 3′ intron termini (the 5′ and 3′ splice sites) and at an internal position near the 3′ end of the intron (the branch site). Although eukaryotic pre-mRNA intron sequences are not conserved, nearly all of them begin with GT and end with AG (7), which, together with their immediate flanking regions, constitute the primary recognition and binding sites for the assembly of spliceosomal components. However, a minor class of nuclear pre-mRNA introns (<0.1%), beginning with AT and ending with AC, have been described in metazoan only (7). These introns contain different splice site and branch point sequences, and are excised by a distinct, low-abundance spliceosome, the 'U12-dependent spliceosome'. Their splicing also requires five snRNAs, U5 being common to both spliceosome types, whereas U11, U12, U4-like and U6-like snRNAs are functionally analogous to U1, U2, U4 and U6-snRNAs, respectively (7). Regardless of the spliceosome type, the splicing reaction itself proceeds via two consecutive *trans*-esterification reactions. In the first reaction, the 5′ splice site is cleaved and the intron 5′ end is ligated to the branch site

*To whom correspondence should be addressed. Tel: +33 1 30 81 54 78; Fax: +33 1 30 81 53 57; Email: claude@grignon.inra.fr

concomitantly. In the second reaction, cleavage of the 3′ splice site releases the intron as a lariat structure, and the 5′ and 3′ exons are ligated.

Although pre-mRNA splicing mechanisms seem to be highly conserved throughout evolution, the exon/intron architecture of eukaryotic nuclear genes exhibits broad interspecies variations. First, spliceosomal introns tend to be larger in multicellular organisms (on average 3000 nt in humans, 1000 nt in rats and mice, 500 nt in chickens, flies and nematodes, and 300 nt in plants) than in unicellular organisms (on average 81 nt in the fission yeast and 21 nt in protozoans) (8,9). Second, the position of introns in protein-encoding genes tends to be conserved in homologous genes in organisms that diverged less than a few hundred million years ago, but not in organisms that diverged 500–1500 million years ago. Third, spliceosomal introns are not distributed with the same frequency in each eukaryotic phylum. They are most common (6 or 7 introns per kb of nuclear genes) in genes from vertebrates and angiosperms. Intermediate densities are observed in other animals and plants, most fungi, slime moulds and ciliates, and very low densities (0.03 introns per kb) are found in some protists (10). These differences raised questions about the evolutionary origin of spliceosomal introns.

Analysis of the complete *Saccharomyces cerevisiae* genome sequence, released in 1996, provided the first step towards understanding the evolution of exon/intron architecture of nuclear genes in yeasts (11,12). It revealed that the nuclear genes of *S.cerevisiae* have an unusual exon/intron architectural organisation compared with other fungi. Only 4% of the *S.cerevisiae* nuclear genes contain introns. Most of these genes (~95%) carry a single medium-sized intron, 300 nt long on average, that is usually located at the 5′ end of the gene, and contains relatively well conserved splice sites. This is in complete contradiction to the exon/intron organisation reported in another yeast species, *Schizosaccharomyces pombe*, in which ~43% of the nuclear genes contain introns, most of which are multiple (up to 15), short (on average 81 nt) and scattered throughout the gene with poorly conserved splice sites (8,13,14).

The significance of the peculiar *S.cerevisiae* exon/intron architecture remains unclear, as only limited information is available from other hemiascomycetous yeasts. Preliminary reports from the *Génolevures* project (15), a comparative sequencing project of hemiascomycetous genomes, have suggested that the *S.cerevisiae* exon/intron architectural organisation is not unique, but is common in other hemiascomycetous yeasts (16–19).

Based on these observations, we decided to investigate the situation further in seven hemiascomycetous yeast species selected across the hemiascomycete tree. Based on a visual detection of spliceosomal introns, we investigated whether certain features (e.g. splicing motifs, position, size, etc.) were conserved among this eukaryotic class, and whether they could potentially serve as markers of nuclear gene evolution. At a higher level, this type of comparative approach may help to understand better how spliceosomal introns are conserved, lost or gained during evolution.

MATERIALS AND METHODS

Source of sequences

The sequences of the seven hemiascomycetous yeasts listed in Table 1 are available at http://cbi.labri.u-bordeaux.fr/Genolevures/Genolevures.php3. Each of the seven yeast genomic libraries contains about 2500 and 5000 random sequence tags (RSTs), which were annotated by using BLAST to compare them with the *S.cerevisiae* proteome plus a set of non-redundant totally sequenced genomes (15).

The *S.cerevisiae* snRNAs sequences used as baits to screen the *Génolevures* RST database are listed in Table 2. The human SR protein SFRS2 (GenBank no. NP_003007) and the *S.pombe* SRP1 protein (GenBank no. AAC49909), a probable metazoan SR-like protein (20), were obtained from NCBI.

Detection of introns in homologous genes

Search for S.cerevisiae-like introns. All RSTs identified by BLAST comparisons as containing *S.cerevisiae* homologues were compared to an updated and non-redundant catalogue of *S.cerevisiae* genes containing intron(s) (see MIPS website, http://mips.gsf.de/proj/yeast/reviews/intron). The comparison generated a set of homologues that may or may not contain intronic structures like their *S.cerevisiae* counterparts. The different splice sites listed in Table 3 were then used as baits to screen this set of sequences, and to determine whether each intron was really present.

Search for introns restricted to a single yeast species. The results of the BLAST search were visually examined to detect any abrupt stop in alignments and any gap between two conserved amino acid sequences of a given protein. This

Table 1. The seven hemiascomycetous yeast species analysed in the *Génolevures* project

Yeast	Strains	Code[a]	RST[b]	Genomic coverage	*Saccharomyces cerevisiae* homologues[c]
Saccharomyces servazzii	CBS 4311	AT	2570	0.2×	1410
Saccharomyces kluyveri	CBS 3082	AU	2528	0.2×	1406
Kluyveromyces marxianus	CBS 712	AZ	2493	0.2×	1301
Pichia angusta	CBS 4732	BB	5082	0.4×	2320
Debaryomyces hansenii	CBS 767	BC	2830	0.2×	1119
Candida tropicalis	CBS 94	BD	2541	0.2×	1130
Yarrowia lipolytica	CLIB 89	AW	4940	0.4×	1078

[a]Code recovered at the beginning of each sequence.
[b]Number of RSTs.
[c]Number of *S.cerevisiae* homologues analysed.

Table 2. *Saccharomyces cerevisiae* homologues of metazoan snRNAs used as baits to screen the different yeast databases

snRNA	Locus	Gene acc no.[a]	Transcript length (nt)	Base-pairing box[b]	Target[c]
U1	*SNR19*	M17205	568	1-AT**ACTTACC**T-10	5′ SS
U2	*SNR20*	M14625	1171	32-**GTGTAGTA**-39	BP
U4	*SNR14*	M17238	160	57-**TGCTGGTT**-64	U6
U6	*SNR6*	X2565	112	55-**GATTAGCA**-62	U4
U6	*SNR6*	X2565	112	43-CAA**TACAGAGA**-53	5′ SS
U5	*SNR7*	M16510	215 (long)-180 (short)	93 –GCCTTTTAC-101	3′ SS

[a]GenBank accession number.
[b]The snRNAs residues involved in the base pairing between snRNAs and intron target sites are shown in bold.
[c]5′SS, BP, and 3′SS indicate the 5′ splice site, the branch site and the 3′ splice site, respectively.

Table 3. Known splice sites used as baits to screen the hemiascomycetous yeast genomic DNA libraries

Yeast species	Number of genes with introns	5′ Motif	Branch motif	3′ Motif
Saccharomyces cerevisiae[a]	*n′ = 252 genes* n = 260 introns	GCAAGT GTATGC GTAAGT GTACGT GTAGTA GTATGT GCATGT GTATGA GTCAGT GTTAAG GTTCGT	AAC TAAC AAT TAAC CAC TAAC GAC TAAC TAC TAAC TACTAAT TAT TAAC TGC TAAC	AAG CAG TAG
Yarrowia lipolytica[b]	*n′ = 13 genes* n = 17 introns	GTAAGT GTGAGT GTGGGT	AAC TAAC ATC TAAC CGC TAAC TAC TAAC	CAG TAG

[a]MIPS database (July 2002 release).
[b]Compilation of intron sequences available at NCBI (February 2002 release).

empirical procedure helped us to identify species-specific introns, which were then validated as described above.

Intron analyses. As RSTs are only 1 kb long on average, only part of the ORF and intron are accessible sometimes. The position of the intron was always obtained from nucleotide sequence alignments derived from protein alignments, therefore we defined intron coordinates according to the nucleotide sequence of the homologue in *S.cerevisiae* (or in *Candida albicans* or *S.pombe* when indicated).

Bioinformatic tools

BLAST (21) was used to screen sequence databases for homology. Sequences were analysed with various programs from the GCG package (Genetics Computer Group, Madison, USA), including 'FINDPATTERNS', a program that detects splicing sites in DNA sequences, 'MAP', a function allowing the detection of ORFs, and FASTA (22). Sequence alignments were generated using CLUSTALW (23) and CLUSTALX (24), and manually adjusted in Genedoc (http://www.psc.edu/biomed/genedoc) to correct obvious mispairings. Phylogenetic trees were generated by the neighbour-joining method (25) and visualised with TreeView, version 1.6.5 (26).

Statistical tools

Chi-square (χ^2) analyses. χ^2 statistical tests were used to assess the statistical significance of the distribution of the splice site motif sequences (5′ site, branch site, 3′ site) relative to the number of introns found in each yeast species. The distribution of reference used was that of *S.cerevisiae*. The 5′ and branching motifs absent from *S.cerevisiae* but detected in other hemiascomycetous yeast species were not taken into account in this analysis, leaving 12 and 10 classes, respectively, for the test.

ANOVA analysis. This statistical test was performed using the SAS system (version 6) to assess whether significant differences existed between *S.cerevisiae* and the other hemiascomycetous yeast species with regard to (i) the size of introns residing in ribosomal and in non-ribosomal protein coding genes, and (ii) the internal distance between the branching and the 3′ splice site. The level of significance accepted was $P = 0.05$.

RESULTS

Frequency of introns in hemiascomycetous yeast genes

A batch of 9764 full length or partial coding sequences, from seven yeast species selected across the evolutionary tree of hemiascomycetes (Table 1) was screened for the presence (or absence) of introns.

We identified 153 partial or complete novel introns: 23 in *Saccharomyces servazzii*, 36 in *Saccharomyces kluyveri*, 19 in *Kluyveromyces marxianus*, 12 in *Candida tropicalis*, 14 in

Table 4. CYGD intron inventory

Spliceosomal intron types		*S. cerevisiae*	*S. servazzii*	*S. kluyveri*	*K. marxianus*	*C. tropicalis*	*D. hansenii*	*P. angusta*	*Y. lipolytica*
***S. cerevisiae*-like**	**Entire**	260	21	20	13	5	9	25	13
	Partial	-	2	9	6	5	1	2	4
	Missing	-	7	0	14	9	6	26	7
***S. pombe* -like**	**Entire**	-	0	1	0	0	0	0	0
New introns	**Entire**	-	0	6	0	2	3	3	2
	Partial	-	0	0	0	0	1	0	0
Number of introns		**260**	**23**	**36**	**19**	**12**	**14**	**30**	**19**
Intron location	**RP** [a]	106	14	10	10	6	5	13	7
	NRP [b]	154	9	26	8	6	9	17	12
	RP/(RP+NRP)	**0.41**	**0.61**	**0.28**	**0.55**	**0.50**	**0.36**	**0.43**	**0.37**
Total number of genes		5261	1410	1406	1301	1130	1119	2320	1078
Number of genes with introns		252	22	36	17	11	12	27	19
Proportion of genes with introns [c]		**4.8**	**1.6**	**2.6**	**1.3**	**1.0**	**1.1**	**1.2**	**1.7**

A complete intron list is available at the MIPS website (http://mips.gsf.de/proj/yeast/reviews/intron/intron_RST.html).
[a]Genes encoding ribosomal proteins.
[b]Genes encoding non-ribosomal proteins.
[c]Expressed in percentages.

Debaryomyces hansenii, 30 in *Pichia angusta* and 19 in *Yarrowia lipolytica* (Table 4). Most of these introns (136 or 88.9%) were located in genes that are known to have an intronic structure in *S.cerevisiae*. Sixteen additional introns were identified in genes that are not known to contain any introns in *S.cerevisiae*. A list of all these introns is available at the MIPS website (http://mips.gsf.de/proj/yeast/reviews/intron/).

The proportion of intron-containing genes appeared to be low in all species studied: from 1.0 up to 2.6%. An estimated 96.8% of these genes contained only one intron. However, some genes contained two introns, like their *S.cerevisiae* counterparts: *S.servazzii* (YCR097c-like), *C.tropicalis* (YER179w-like), *D.hansenii* (YLR367w- and YOL017w-like) and *P.angusta* (YDR424c-, YPL198w- and YLR367w-like). We putatively identified a second intron in the *K.marxianus* YFL034cb-like gene, whereas its *S.cerevisiae* counterpart has only one intron.

Splice sites

Most of the splice sites detected at the exon/intron boundaries exhibited the canonical 5′-GT…AG-3′ sequence (Fig. 1). Only 2.2% of the introns from *S.cerevisiae* and 6.7% of those from *Y.lipolytica* exhibited a non-canonical 5′-GC…AG-3′ sequence at these locations. Thus, it appears that splice sites are well conserved between yeasts, although a certain degree of flexibility exists, as described below.

5′ splice site. The extended GTATGT motif found in *S.cerevisiae* appeared to be more or less flexible depending on the yeast species: 12 different 5′ motifs were found in the *S.cerevisiae* intronic genes, whereas only one was observed in *S.servazzii* and *C.tropicalis*. The χ^2 analysis clearly shows that *S.cerevisiae*, *S.servazzii*, *S.kluyveri*, *K.marxianus* and *C.tropicalis* share the same motif distribution with an over-representation of GTATGT as 5′ motif. In contrast, this distribution was different ($P = 0.001$) in *D.hansenii* in which GTATGT represents only 46.2% of the 5′ splice motifs, and in *P.angusta* and *Y.lipolytica*, in which GTAAGT and GTGAGT are the predominant 5′ splice motifs, respectively. The GTAGGT motif in *P.angusta* was detected several times, and therefore considered to be a novel 5′ splice motif. Some other unusual motifs were also detected but classified as putative 5′ splice motifs, as they were only found once in a given yeast species, and may reflect sequencing errors.

Branch site. Branch motifs appeared to be more strictly conserved. Between 82.9 and 94.7% of the introns contained TACTAAC as branch motif in *S.cerevisiae*, *S.servazzii*, *S.kluyveri*, *K.marxianus* and *C.tropicalis*. This percentage decreased to between 53.3 and 41.4% in the three other yeast species, which are more distantly related to *S.cerevisiae*. Although TACTAAC was the most frequent branch motif, χ^2 tests show that *D.hansenii*, *P.angusta* and *Y.lipolytica* statistically ($P = 0.001$) have a different branch motif distribution than the other yeast species with alternative branch motifs more frequently used. These results parallel the change in the preferred 5′ motif. Two novel branch motifs were also detected: ATCTAAT in *D.hansenii* and *P.angusta*, and TTCTAAC in *C.tropicalis* and *Y.lipolytica*. Other putative

	Motifs	*S. cerevisiae*	*S. servazzii*	*S. kluyveri*	*K. marxianus*	*C. tropicalis*	*D. hansenii*	*P. angusta*	*Y. lipolytica*
5' splice site	**GTATGT**	**198**	**22**	**21**	**11**	**7**	**6**	4	1
	GTACGT	24	-	3	-	-	-	1	-
	GTATGA	10	-	-	-	-	-	-	-
	GTAAGT	9	-	2	-	-	3	**19**	1
	GTATGC	9	-	2	-	-	-	-	-
	GCATGT	4	-	-	-	-	-	-	1
	GTGAGT	1	-	-	-	-	-	-	**12**
	GTCAGT	1	-	-	-	-	-	-	-
	GTTCGT	1	-	-	-	-	-	-	-
	GTAGTA	1	-	-	-	-	-	-	-
	GTTAAG	1	-	-	-	-	-	-	-
	GCAAGT	**1**	-	-	-	-	2	-	-
	GTAGGT	0	-	-	-	-	-	4	-
	GTGAAA	0	-	-	-	-	-	-	1
	GTACGA	0	-	-	1	-	-	-	-
	GTATAC	0	-	-	-	-	1	-	-
	GTACGG	0	-	-	1	-	-	-	-
	GTGGAT	0	-	-	-	-	1	-	-
	Sampling	**n=260**	**n=22**	**n=28**	**n=13**	**n=7**	**n=13**	**n=28**	**n=16**
	P-value						0.001	0.001	0.001
Branch site	**TACTAAC**	**226**	**20**	**29**	**18**	**11**	**8**	**12**	**8**
	GACTAAC	12	1	-	-	-	-	-	-
	AACTAAC	7	1	3	1	-	2	2	4
	CACTAAC	5	-	-	-	-	-	2	1
	TGCTAAC	3	-	1	-	-	2	3	2
	TATTAAC	2	-	1	-	-	1	-	-
	AATTAAC	2	-	-	-	-	1	1	-
	ATATAAC	1	-	-	-	-	-	-	-
	GAATAAC	1	-	-	-	-	-	-	-
	TACTAAT	1	-	-	-	-	-	-	-
	TTCTAAC	0	-	-	-	1	-	-	1
	ATCTAAT	0	-	-	-	-	1	1	-
	CGCTAAC	0	-	-	-	-	-	1	2
	TCAAAAC	0	-	1	-	-	-	-	-
	AACTAAT	0	-	-	-	-	-	1	-
	ATCTAAC	0	-	-	-	-	-	1	-
	TGCTAAT	0	-	-	-	-	-	1	-
	CTCTAAC	0	-	-	-	-	-	1	-
	GTCTAAC	0	-	-	-	-	-	1	-
	GCCTAAC	0	-	-	-	-	-	1	-
	AGCTAAC	0	-	-	-	-	-	1	-
	Sampling	**n=260**	**n=22**	**n=35**	**n=19**	**n=12**	**n=15**	**n=29**	**n=18**
	P-value						0.001	0.001	0.001
3' splice site	**TAG**	**137**	7	10	**7**	**11**	**11**	**17**	2
	CAG	**116**	**13**	**23**	**10**	-	**3**	**11**	**15**
	AAG	7	2	2	2	1	1	2	1
	Sampling	**n=260**	**n=22**	**n=35**	**n=19**	**n=12**	**n=15**	**n=30**	**n=18**
	P-value		0.05	0.05		0.01			0.01

Figure 1. Compilation of splice motifs. The different types of 5′ splice motifs, branch motifs and 3′ splice motifs found are listed on the left, and the name of the yeast species is on the top. The occurrence of each splice motif is indicated for each species. 'n' Represents the total number of motifs analysed in each species. The level of significance of the χ^2 statistical tests is indicated by the P-value. The character '-' indicates that the corresponding motif was not found in our set of intron sequences. The preferential motifs are in indicated in bold, the confirmed novel motifs (encountered several times) are underlined, and putative novel splice motifs (encountered once only) are in italics.

motifs were also found, but require as above additional confirmation.

3′ splice site. Only three 3′ splice motifs exist in *S.cerevisiae*, YAG (Y = pyrimidine) being more frequent than AAG. Like in *S.cerevisiae*, TAG and CAG were shown to be equally frequent (~50%) in *K.marxianus*, *D.hansenii* and *P.angusta*. In contrast, *S.servazzii* and *S.kluyveri* ($P = 0.05$) as *Y.lipolytica* ($P = 0.01$) were found to preferentially use CAG as 3′ motif, whereas *C.tropicalis* clearly uses TAG as preferential 3′ motif,

suggesting that the distribution of the 3′ splice motif is unrelated to the phylogenetic position of the yeast species unlike the 5′ and the branch motifs.

Non-random vocabulary within exon and intron sequences

A BLAST search was carried out to compare *S.cerevisiae* intron sequences with their homologues in the other hemiascomycetous yeasts (data not shown). As expected, no significant long stretches of conserved sequences were observed in introns at identical positions. However, a non-random vocabulary was observed at critical sites within the intron sequence.

A bias in AT residues. In all of the yeast species studied, introns tended to be slightly richer in A+T residues than did their neighbouring exons (from 2.9 to 7.8%) (data not shown, see the MIPS website for more details).

A bias in nucleotide composition around splice sites. The GC content of exons and intron sequences was taken into account to evaluate the over-representation of some residues around splice sites in each yeast species. Briefly, nucleotides were considered to be preferentially conserved at a given position when their individual proportion represented at least 40% of the nucleotides registered at this location. The conservation of the exonic nucleotides adjacent to the splice sites was analysed (Fig. 2). Thus, the conserved sequence was: AAG|NNN for *S.cerevisiae*, A(A/T)G|TTA for *S.servazzii*, AAN|NNN for *S.kluyveri*, A(A/C)G|GCT for *K.marxianus*, A(A/C)G|TCT for *C.tropicalis*, (A/T)A(A/T)|NAA for *D.hansenii*, ANG|GCT for *P.angusta* and NN(A/G)|NAA for *Y.lipolytica* (the slash indicates the exon boundaries). Thus G generally precedes the 5′ splice site in most yeast species. This nucleotide preference can be enlarged to AAG in the yeast species most closely related to *S.cerevisiae*, the only clear exception being *D.hansenii* in which (A/T)A(A/T) is the consensus detected at the 5′ exon/intron boundary. The intronic context of the splicing sites also appeared to be somewhat conserved. Except for a few exceptions, an A and a T were always found in the downstream region of the 5′ site, around the branch site and in the upstream region of the 3′ site. All of these observations helped us to define an extension of the core splice motifs in each yeast species (see the MIPS website).

Intron size

Mean intron size. *Saccharomyces cerevisiae* (263.98 nt), *S.servazzii* (277.2 nt) and *K.marxianus* (278.5 nt) appeared to have large introns, *C.tropicalis* (192.0 nt) has a moderate intron size, whereas the most distant yeasts species *D.hansenii* and *P.angusta* have short introns (129.5 and 92.8 nt, respectively) (Fig. 3A). Two inconsistencies were found with *S.kluyveri* having a moderate intron size (158.2 nt) whereas a large one is expected, and with *Y.lipolytica* having large introns (303.9 nt) whereas short ones are expected. The moderate mean intron size found in *S.kluyveri* is probably under-estimated due to library biases. Most *S.kluyveri* introns were indeed identified in non-ribosomal protein (NRP) genes that contain smaller introns (see below). However, it is difficult to compare mean intron size as high standard deviations were observed in all seven yeast species, partly due to the fact that introns residing in ribosomal protein (RP) genes are larger than those residing in NRP genes (Fig. 3A). Indeed, as usually occurs in *S.cerevisiae* (27), the distribution of intron lengths was found to follow a bimodal rule in the seven other yeast species, with a first peak at ~50–100 nt, corresponding to introns in NRP genes, and a second peak near 250–400 nt, corresponding to introns in RP genes (see the MIPS website). When the two groups were analysed separately, it appears that the mean size of RP introns from the *Saccharomyces–Kluyveromyces* clade (between 300.3 and 398.6 nt) was statistically different ($P = 0.05$) from that of the less closely related yeasts like *C.tropicalis* (290.2 nt), *D.hansenii* (255.2 nt) and *P.angusta* (112.2 nt). This suggests a size variation in direct connection with the phylogenetic position of the yeasts. *Yarrowia lipolytica* stands apart with a mean RP intron size similar to that encountered in *S.cerevisiae*. No clear picture was found when the mean size of NRP introns was analysed, except that *Y.lipolytica* NRP introns are significantly larger than those of *S.cerevisiae*, whereas *S.kluyveri* and *P.angusta* NRP introns are smaller. It thus remains difficult to relate mean intron size to phylogenetic position.

Internal distances between splicing sites. Unlike the distance between the 5′ site and the branch site (S1), which was found to vary widely in size (e.g. from 23 up to 937 nt) according to a bimodal distribution, the distance between the branch site and the 3′ site (S2) varied over a narrower range, and displayed a unimodal distribution (see MIPS). The mean length of S2 appeared to be similar between RP and NRP introns in each yeast species, but varies according to the phylogenetic position of the yeast species (Fig. 3B). Its length decreases from 45.8 and 33.2 nt in the *Saccharomyces–Kluyveromyces* clade (except in *S.kluyveri*) to 1.8 nt in *Y.lipolytica* (Fig. 3B). This suggests that the S2 internal distance may be an evolutionary marker in hemiascomycetous yeasts.

Intron location

About 93% (i.e. 93.5%) of the introns detected were located inside an ORF (see MIPS). Most of them (87.0%) were near to the 5′ end or within the first 100 nt (most often the first 25 nt) following the start codon. Less than 7% (i.e. 6.5%) of the introns were located outside of ORFs, in the 5′ untranslated regions (UTRs). No introns in the 3′ UTR have been detected. The proportion of these ‘external’ introns was probably under-estimated due to our methodology of detection.

Conservation of introns in paralogous and homologous genes

Intron presence/absence. The presence or absence of intronic structures was compared in *S.cerevisiae* paralogues and homologues (Fig. 4). Nearly 58% of the *S.cerevisiae* introns were located in paralogous genes distributed in 85 distinct gene families. An intron was detected in all members of 43 of the 85 families (i.e. 50.6%). In contrast, 42 gene families of 1 to 10 members (49.4%) did not contain any introns. Similar trends were observed on homologues in other species. Some introns that were absent from *S.cerevisiae* genes were detected in homologous genes (group A), and some that were present in *S.cerevisiae* were not detected in the homologues (group B). Several introns present in *S.cerevisiae* and in the most

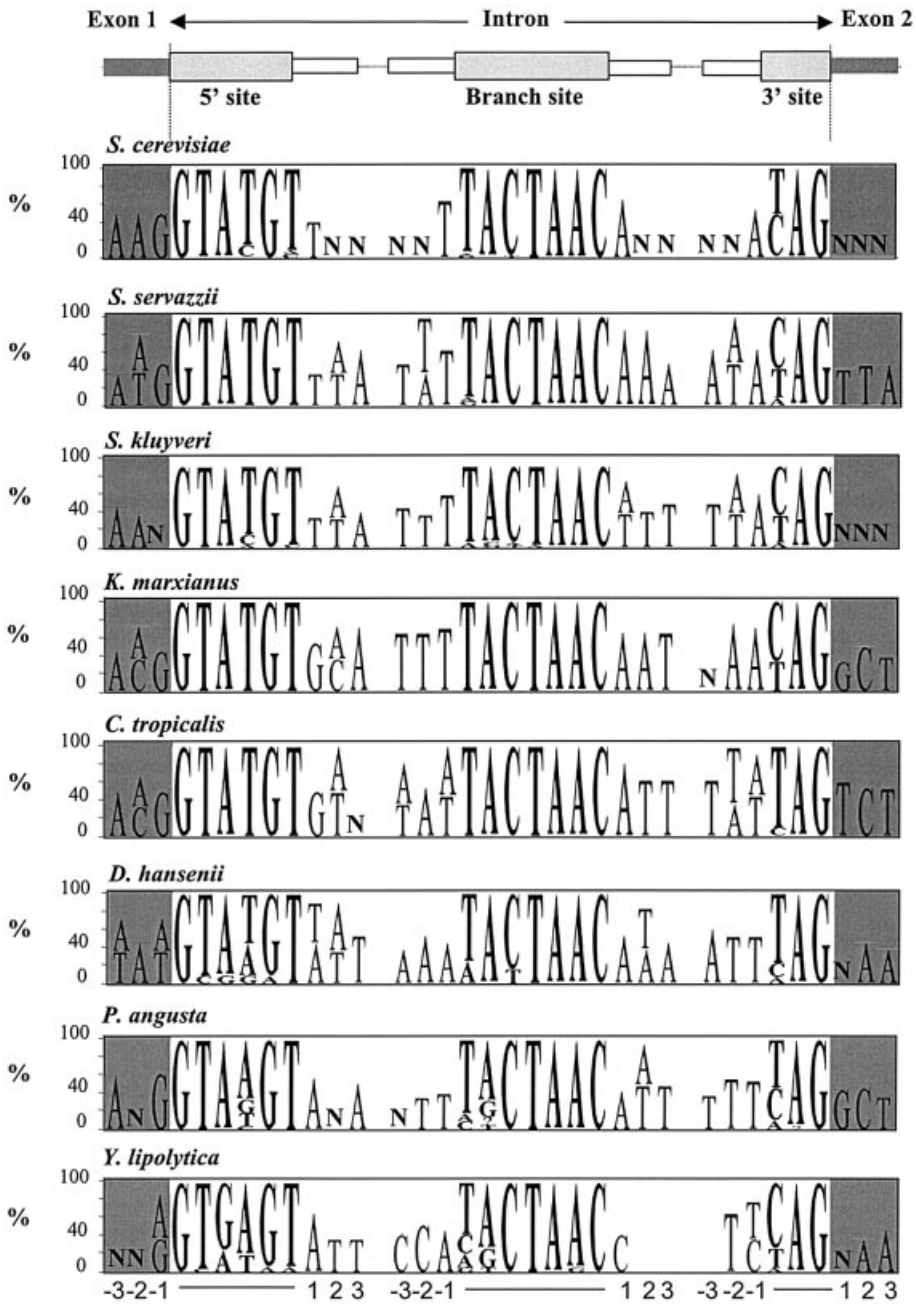
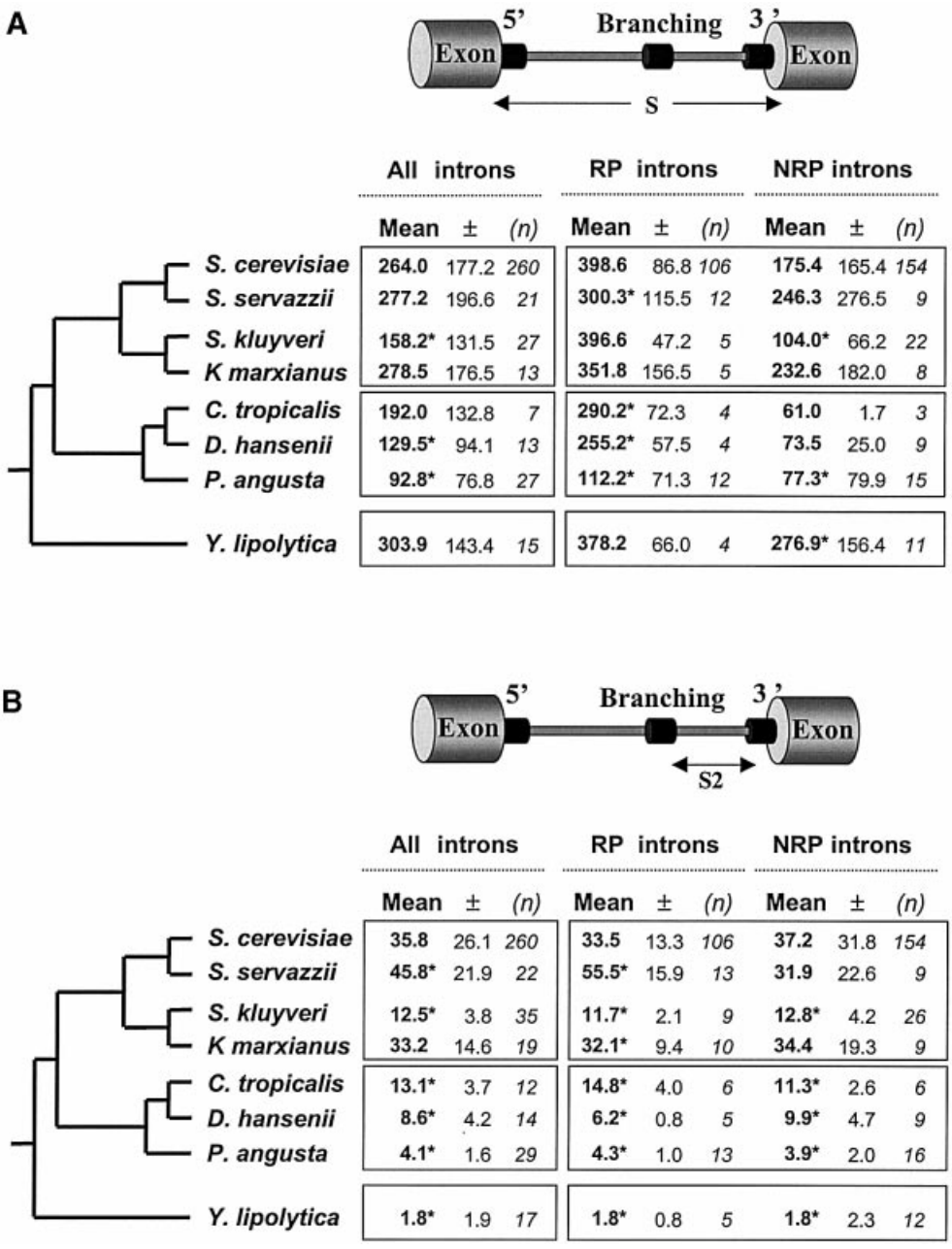


Figure 2. Preferential nucleotide usage around splicing sites. The vertical and horizontal axes represent the frequencies and positions of nucleotides, respectively. The height (expressed as a percentage) of each letter is proportional to the frequency of each base at each position.

distantly related species were absent from the intermediate yeast species (group C).

Conservation of intron position. Figure 5 presents some examples of the tracking of intron position between homologous yeast genes. Intron position was defined at the nucleotide level making it possible to determine whether the intron was at an identical position (same codon and same phase) in paralogous or orthologous genes. CLUSTALX comparisons of exon/intron structures were used to reconstruct a virtual (scaffold) gene, exhibiting the probable ancestral intron arrangement. These multiple alignments suggested that the positions of intronic structures are always highly conserved when they are located within an ORF (Fig. 5A). When an intronic structure was absent from one or several homologous genes (Fig. 5B), we observed that the 3′ introns were preferentially affected. Thus, the YFL037w scaffold gene contains four found introns, which are either present or absent in the different homologues. Introns located near the 5′ end, like i1 and i2, are present in *C.albicans* and *S.pombe*. Only i1 is conserved in *S.kluyveri*, whereas the *S.cerevisiae* homologue is completely devoid of introns. Some examples in which the intron position at the 5′ end of the protein-coding gene sequence was not conserved were also found, but these



	All introns			RP introns			NRP introns		
	Mean	±	(n)	Mean	±	(n)	Mean	±	(n)
S. cerevisiae	264.0	177.2	260	398.6	86.8	106	175.4	165.4	154
S. servazzii	277.2	196.6	21	300.3*	115.5	12	246.3	276.5	9
S. kluyveri	158.2*	131.5	27	396.6	47.2	5	104.0*	66.2	22
K marxianus	278.5	176.5	13	351.8	156.5	5	232.6	182.0	8
C. tropicalis	192.0	132.8	7	290.2*	72.3	4	61.0	1.7	3
D. hansenii	129.5*	94.1	13	255.2*	57.5	4	73.5	25.0	9
P. angusta	92.8*	76.8	27	112.2*	71.3	12	77.3*	79.9	15
Y. lipolytica	303.9	143.4	15	378.2	66.0	4	276.9*	156.4	11

	All introns			RP introns			NRP introns		
	Mean	±	(n)	Mean	±	(n)	Mean	±	(n)
S. cerevisiae	35.8	26.1	260	33.5	13.3	106	37.2	31.8	154
S. servazzii	45.8*	21.9	22	55.5*	15.9	13	31.9	22.6	9
S. kluyveri	12.5*	3.8	35	11.7*	2.1	9	12.8*	4.2	26
K marxianus	33.2	14.6	19	32.1*	9.4	10	34.4	19.3	9
C. tropicalis	13.1*	3.7	12	14.8*	4.0	6	11.3*	2.6	6
D. hansenii	8.6*	4.2	14	6.2*	0.8	5	9.9*	4.7	9
P. angusta	4.1*	1.6	29	4.3*	1.0	13	3.9*	2.0	16
Y. lipolytica	1.8*	1.9	17	1.8*	0.8	5	1.8*	2.3	12

Figure 3. Yeast spliceosomal intron length. (**A**) Phylogenetic evolution of mean intron size in all introns, and in genes encoding NRPs and RPs. (**B**) Phylogenetic evolution of the mean S2 distance, e.g. the internal distance between the branch motif and the 3′ splice motif in: all introns, in genes encoding NRPs and in those encoding RPs. For each yeast species and each intron population: the mean intron size (Mean), the standard deviation (±) in nucleotides and the number of introns analysed (*n*) are indicated. Panels illustrate the main phylogenetic subdivisions. The asterisk represents that this value was found to be statistically different from that observed in *S.cerevisiae* at $P = 0.05$.

appeared to be exceptional. Thus, the YMR004w scaffold gene exhibits two intron positions of which only i2 is conserved in the *C.albicans* homologue.

Some complex situations were found in which the positions of introns in homologous genes were close but not identical (Fig. 5C). Most of these cases were encountered in homologous genes that shared poor sequence similarity in the region of interest, suggesting a possible artefactual effect. For example, two introns (i1 and i2) were detected in the *C.albicans* YBL099w homologue as well as in its *S.pombe* counterpart (i′1 and i′2). When the position of i1 was compared to that of i′1, it appeared that i1 and i′1 were in the same phase (GTT | GCT for i1; TCT | ACA for i′1; where the slash indicates the position of the intron within the codon), but in a different location in the scaffold gene (Δ = 15 nt). i2 and i′2 were in a different phase (GG|T for i2 and GAG | GAA for i′2) and at a different location (Δ = 65 nt).

Rare cases of divergent intron positions were found among homologous genes that exhibited a very good sequence conservation (>90% identity) in the region of interest, such as is the case for YGR001c and YIL148w (Fig. 5C). In YGR001c, two intron positions exist in the *S.cerevisiae* ORF

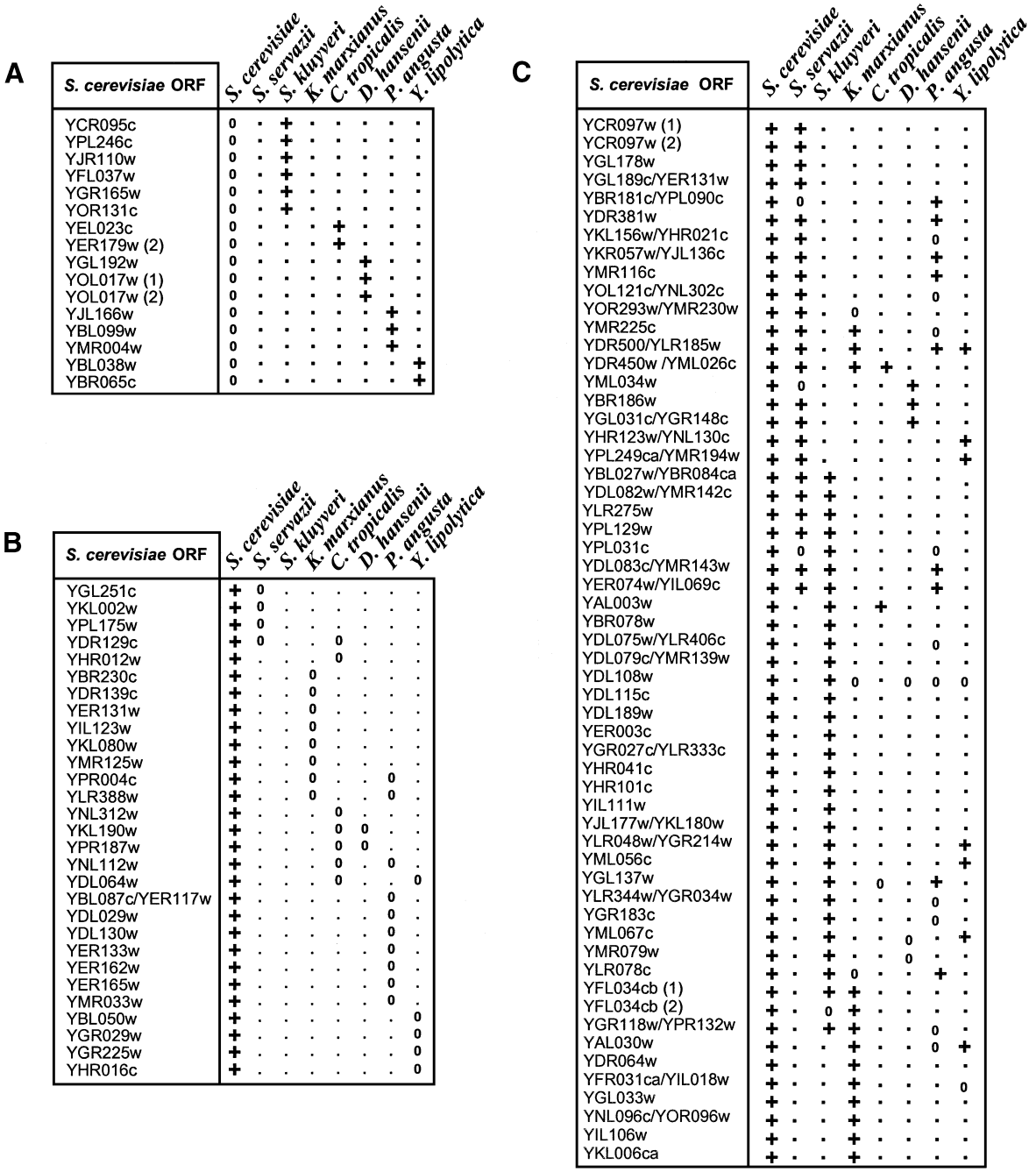


Figure 4. Presence or absence of introns in homologous genes of hemiascomycetous yeasts. A group of 52 introns was shown to be present in at least three yeast species. This figure indicates whether each intron was present or absent in each yeast species. The name of the *S.cerevisiae* ORF is indicated on the left. The presence of an intron is indicated by '+' and its absence by '0'. The symbol '.' indicates that the homologous gene was not found in our DNA library. Introns apparently restricted to a yeast species other than *S.cerevisiae* are in (**A**), those restricted to the *S.cerevisiae* species are in (**B**), and those that can be considered to be 'ubiquitous' are in (**C**).

(i1 and i2), whereas only one of these positions is present in the *Y.lipolytica* counterpart (i′1). When i1 was compared with i′1, both introns appeared to interrupt the ORF at the same position within the codon (TT|G for i1 and CA|A for i′1), i.e. the intron phase is the same, but not at the same location (Δ = 3 nt) within the scaffold gene. In YIL148w, a single intron is present in the *S.cerevisiae* ORF (i1) and in its *Y.lipolytica* counterpart (i′1), but they are not located at identical positions (Δ = 7 nt) and their phases are different (AT|T for i1 and A|TG for i′1). In both cases, we cannot exclude that the same intron migrated to a nearby location, or that two ancestral introns coexisted, separated by 3 or 7 nt. Additional examples are available at the MIPS intron website.

Conservation of spliceosomal components

In order to check whether splicing site variability might be correlated with variation of the splicing machinery, we

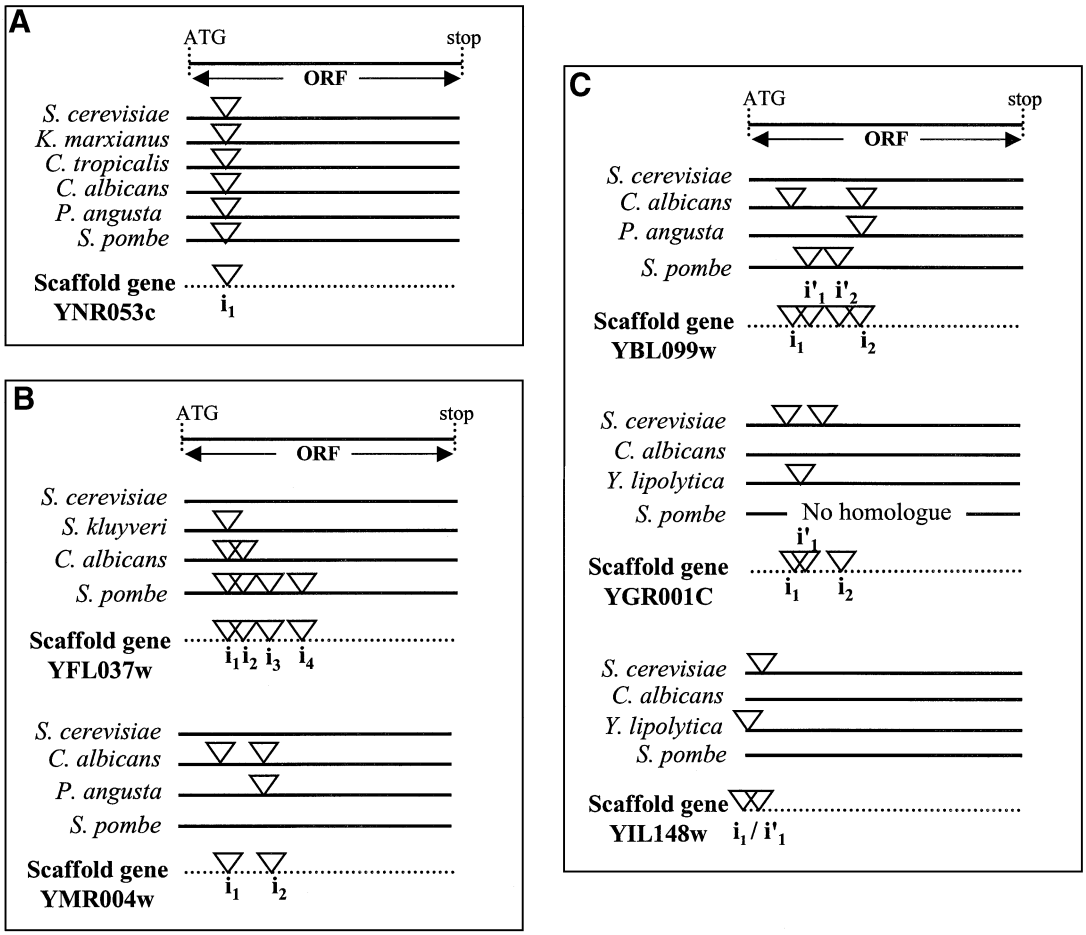


Figure 5. Conservation of intron positions. Examples of the tracking of intron positions between homologous yeast genes. (**A**) Conserved intron positions. (**B**) Missing intron positions at the 3′ end (YFL037w) or at the 5′ end (YMR004w) of the gene sequences. (**C**) Divergent intron positions in poorly (YBL099w) or highly (YGR001c, YIL148w) conserved regions. The horizontal lines illustrate the different sequences aligned by ClustalX. Intron positions are indicated by the inverted triangle symbol. A scaffold gene is indicated by dashed lines in each example.

searched for two types of components: snRNAs and SR-splicing factors.

Conservation of snRNAs. Homologues of metazoan U1, U2, U4, U5 and U6 snRNAs were found in hemiascomycetous yeast species, including *C.albicans* (see the MIPS website) on the basis of sequence conservation and shared structural elements. The functional features like the intron base-pairing site and the Sm site ($PuAT_{4-6}G$-Pu), involved in the contact between the snRNAs and the splice sites and the Sm proteins, respectively (28), were systematically identified when possible. Previous work (28,29) has reported similar findings in 24 yeasts including *S.cerevisiae* and some of the yeast species herein studied: *S.servazzii*, *S.kluyveri*, *Kluyveromyces lactis* and *Y.lipolytica*. Our results complete this original survey and identifies snRNAs homologues in *Zygosaccharomyces rouxii* (U1, U5), *Kluyveromyces thermotolerans* (U1, U6), *K.marxianus* (U5), *C.albicans* (U2, U4, U5, U6), *D.hansenii* (U4), *P.angusta* (U5, U6) and *Pichia sorbitophila* (U6) (see MIPS). All functional features exhibit a nearly perfect sequence conservation when each yeast snRNA homologues were compared with their metazoan counterparts. The analyses of the size and general organisation of each snRNA molecule corroborates the findings of previous works (28,29). (i) U1, U2 and U5 snRNA homologues are highly polymorphic in size and general organisation in the hemiascomycetous yeasts compared with their eukaryotic counterparts. The comparison of their sequences clearly shows that such discordance is due to the presence of supplementary regions that are more or less phylogenetically variable (see MIPS). These 'extra' residues probably provide phylogenetic support for variable secondary structures (28,29), but their precise function remains to be clarified. (ii) U4 and U6 snRNA homologues are relatively well conserved in length and sequence (see MIPS), probably reflecting evolutionary constraints imposed by the interaction between these molecules (28) as evidenced by the high conservation of the U6 base-pairing site sequence 'TGCTPuPuTT' in the U4 snRNA homologues, and conversely that of the U4 base-pairing site sequence 'PuATPyAGCA' in the U6 snRNAs homologues (see MIPS).

SR-splicing factors. Putative homologues to mammalian SR-like splicing factors were detected in *S.kluyveri*, *K.lactis*,

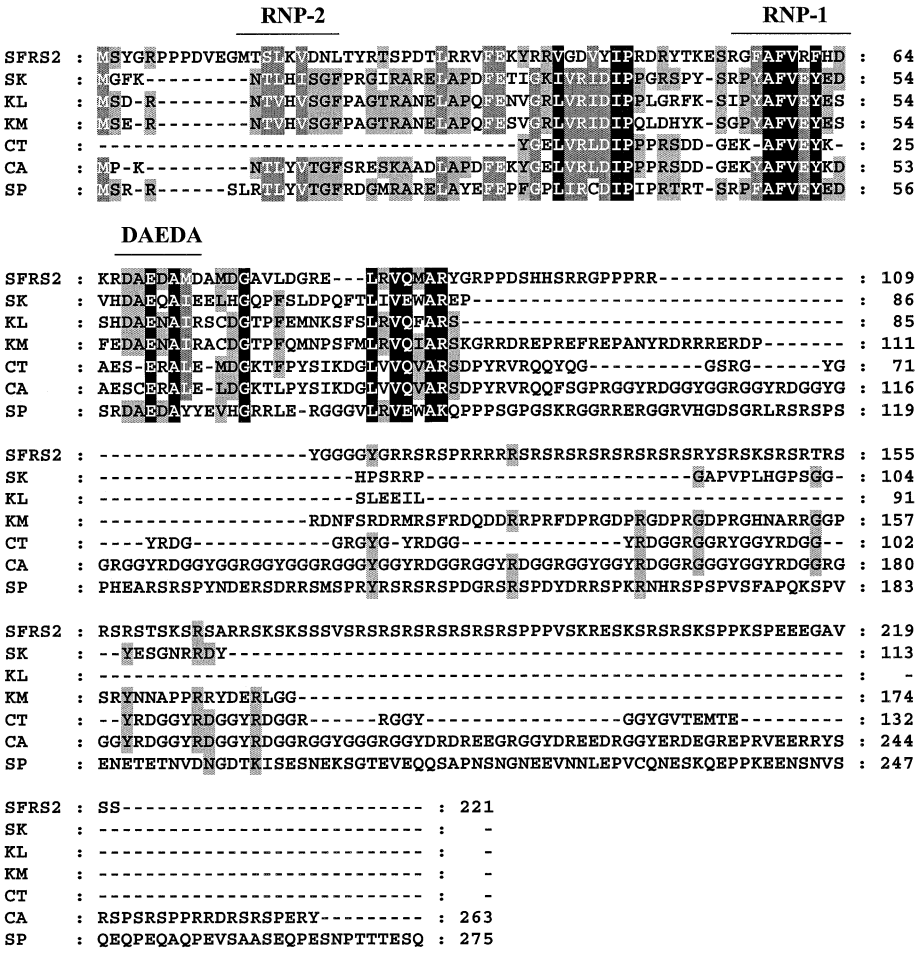


Figure 6. Putative homologues of metazoan SR-splicing factors. ClustalX multiple alignments of the deduced amino acid sequences of the newly identified putative hemiascomycetous SR-like splicing factors in *S.kluyveri* (SK), *K.lactis* (KL), *K.marxianus* (KM), *C.tropicalis* (CT) and *C.albicans* (CA), with the human SFRS2 and the *S.pombe* (SP) Srp1 amino acid sequences. The positions of RNP-2, RNP-1 and DAEDA domains are indicated. The positions of conserved residues are highlighted by vertical shading.

K.marxianus, *C.tropicalis* and *C.albicans* (Fig. 6). In contrast, no clear homology was detected in *S.cerevisiae*. When sequences were analysed, it appeared that unlike the *S.pombe* SRP1 gene, which has two introns, the *S.kluyveri*, *C.tropicalis* and *C.albicans* SRP1-like DNA sequences have only one intron. No introns were detected in *K.lactis* or *K.marxianus* homologues. CLUSTAL multiple alignments clearly show that the putative hemiascomycetous SR-like splicing factors share some conserved domains, like their counterparts in humans and in *S.pombe*. The N-terminal region contains the RNP-2, RNP-1 and DAEDA signature sequences, which are characteristic RNA-binding domains. The C-terminal region appeared to be more variable and to consist of alternating dipeptides of various lengths and arrangements in each species: RS (arginine/serine) in human and in *S.pombe*, RD (arginine/aspartic acid) or RG (arginine/glycine) in the other yeasts. This raises the intriguing possibility that, like *S.pombe*, some hemiascomycetous yeasts could have conserved some of the functions required for mRNA alternative splicing (30).

DISCUSSION

As mentioned in the Introduction, *S.cerevisiae* and *S.pombe* differ greatly with respect to the architectural organisation of their nuclear genes, the former being intron-poor and the latter intron-rich with degenerate splicing consensus, like filamentous fungi and most eukaryotic organisms (8). The aim of this study was to determine whether the unusual intron situation observed in *S.cerevisiae* could be extended to the rest of the hemiascomycetous yeasts. This is indeed the case. Our results clearly show that hemiascomycetous spliceosomal introns obey a general rule that sets them apart from other eukaryotes, including most fungi.

Hemiascomycetous yeast genomes are intron poor

Only a very small proportion of hemiascomycetous nuclear genes are interrupted by an intronic structure. In *S.cerevisiae*, 260 spliceosomal introns have been detected in 252 nuclear genes (see the MIPS website), representing only 4.8% of the nuclear genes. This situation is also observed in other yeast

species, in which the proportion of detected intron-containing genes varies between 1.0% in *D.hansenii* and 2.6% in *S.kluyveri*. These low percentages are probably underestimated as analyses were performed on single reads and on partial yeast genomes (0.2× to 0.4× coverage) and as natural introns may have non consensus splice signals or lie in maverick genes, i.e. genes with no homologues or questionable ORFs (31), thus escaping detection. Nevertheless, hemiascomycetous yeasts appear to be different from other fungi like *Neurospora crassa* and *S.pombe*, and also from plants and vertebrates, in which intronless genes are exceptional (8,14).

This paucity of intron-containing genes raises several questions. (i) If ancient ascomycetes like *S.pombe* had many introns, this suggests that introns were massively erased from nearly 95% of hemiascomycetous genes. This point will be discussed further below. (ii) A related question concerns the small number of genes that have retained introns. In these cases, the surviving introns may confer a selective advantage to their host. In *S.cerevisiae*, mRNAs corresponding to intron-containing genes account for ~26% of all mRNA transcripts, and 90% of them are derived from genes coding for RPs (11,32). The elevated proportion of introns residing in RP genes, i.e. from 41% in *S.cerevisiae* and up to 61% in *S.servazzii*, suggests that introns are somehow required in hemiascomycetous yeasts to maintain and to control high ribosome biogenesis as they apparently do in mice (33) and in *S.cerevisiae* (34). As this correlation is negative in multicellular organisms and absent in some others (34), this hypothesis needs to be investigated further.

One possible evolutionary scenario is that the scarcity of intron-containing genes favoured a reduction in genome size (35) (typically <20 Mb in yeasts) and the shortening of the S-phase, therefore leading to adaptation to a specific life style (15,36). This is consistent with the notion that hemiascomycetes are rapidly dividing unicellular organisms, whereas higher eukaryotes divide rather slowly (37,38). However, this does not explain why hemiascomycetes are intron poor and why *S.pombe* is intron rich, as all these yeasts have similar sized genomes. One likely hypothesis to explain such discrepancies involves lineage-specific loss and divergence. Nearly 200 genes have been lost or have diverged since the radiation of the *S.cerevisiae* lineage from the last common ancestor with *S.pombe*. Most of these genes were involved in splicing and in the posttranscriptional gene silencing system (PTGS) (39). As one of the functions of this PTGS system is to 'tame' retrotransposons, the loss of this system in the *S.cerevisiae* lineage may have induced a burst of retrotransposition, erasing most of the introns from the nuclear genes (39).

Introns are asymmetrically distributed along the protein-coding gene

Ninety-five percent of intron-containing hemiascomycetous genes are mono-intronic and nearly all introns (87.0%) are located in the first few codons of the 5′ end. This is not consistent with the general picture observed in most eukaryotic organisms (8), notably *S.pombe* in which nuclear genes are generally multi-intronic with introns scattered all along the length of ORFs (14). However, one-third of *S.pombe* genes have an intron at their 5′ end (13), suggesting that such 5′ bias in intron distribution also exists in *S.pombe* (14), even though it is less pronounced than in the hemiascomycetous yeasts.

We also noticed that the position of a given intron is generally conserved in paralogues and orthologues. This supports the hypothesis that intron positions are critical for the efficient splicing efficiency of pre-mRNA (40). However, cases of intron 'disappearance' were also observed in a non-negligible proportion of homologous yeast genes. We showed that when one of several introns is missing from a homologous gene, it is usually the most 3′ one that is missing. The model proposed by Fink (38) provides a plausible explanation for this asymmetric distribution. It speculates that yeast genes lost their introns due to homologous recombination with reverse-transcribed cDNAs initiated at the 3′ end of the message. Non-LTR retrotransposable elements (LINEs) might be the source of such endogenous RT activity, which produces pseudogenes in humans, unlike retroviruses (41,42). LINE elements were recently detected in *C.albicans* (43) and *Y.lipolytica* (44), but have not been found in other hemiascomycetous yeasts. As 50% of the *S.cerevisiae* paralogous genes have lost several introns even though they were derived from recent duplication events, LINE-independent intron removal may exist. To date, intron loss resulting from retrotransposition has only been reported in *Arabidopsis thaliana* (45).

Mean intron sizes can vary widely

Mean intron size is not uniform among hemiascomycetes. In fact, it varies according to the phylogenetic position of the yeast species. It is 265.5 nt in *Saccharomyces* and *Kluyveromyces* species, which is close to the size of vertebrate and plant introns (8). However, it tends to decrease progressively in the more distant yeast species, reaching 92.8 nt in *P.angusta*, which is in a size range comparable to that in *S.pombe* (14) and most fungi. The mean intron size varies widely, however, reflecting the existence of two distinct populations of introns (27) (see MIPS): short introns (50–100 nt long) are preferentially located in NRP genes, and large introns (>250 nt long) tend to be found in RP genes.

We found that the mean distance between the branch site and the 3′ splice site correlates with the phylogenetic position of the yeast species, ranging from 30 to 40 nt in the *Saccharomyces* and the *Kluyveromyces* species (except perhaps *S.kluyveri*) to 2 nt in *Y.lipolytica* (the most distant species). This feature may be a more reliable evolutionary marker than the mean intron size, at least in the hemiascomycetous yeasts.

All introns share a non-random vocabulary

Nearly all hemiascomycetous yeast introns begin with 'GT' and end with 'AG', whereas introns beginning with 'GC' are exceptional with only five examples in *S.cerevisiae* (11), like in *S.pombe* and other eukaryotes (14). The predominance of this 'GT…AG' rule supports the hypothesis that spliceosomal introns are spliced in yeasts, as in most eukaryotes, by a conserved mechanism, which correlates the strong conservation of snRNAs observed from mammals to *S.pombe* (28).

In contrast, clear differences were noted in the splice site motifs. They tend to be well conserved in hemiascomycetous yeast and are typically less degenerate than in most eukaryotes, including fungi (3,14). We detected 12 distinct

5′ motifs in *S.cerevisiae* and 17 in the hemiascomycetous yeast species compared with 100 in *S.pombe* (see ftp://ftp.sanger.ac.uk/pub/yeast/pombe/Intron_Data/). However, a preferred 5′ motif exists in each yeast species (GTATGT, GTAAGT or GTGAGT) in spite of the conservation of the base-pairing box of the snRNAs (see MIPS). The functional significance of such predominant 5′ motifs remains to be elucidated. GTAAGT, which perfectly complements the U1 snRNA base-pairing box, may represent the ancestral 5′ motif. Similarly, TACTAAC, which was the most common branch motif, nearly perfectly complements the U2 snRNA base-pairing box. In contrast, only shorter and more degenerate motifs are found at this location in human and *S.pombe*, in which YTRAY is proposed to be the consensual branch motif (46). The 3′ splice site motifs appear much more conserved with only three possible motifs in eukaryotes: TAG, CAG and AAG (2,47,48). YAG was the most common 3′ motif in hemiascomycetes.

Our analysis showed that non-random sequence patterns are conserved at sites flanking the core splicing sites in all yeast introns, which is in agreement with previous findings (11,49). The nature of this preferential nucleotide usage appears to be species dependent although some general rules exist, for example the AAG-intron-G is conserved in many yeast species. The significance of such conservation remains unclear. These non-random sequences closely resemble the 'proto-splice site' (MAG-intron-R) postulated for intron insertion events (1,3,50,51), suggesting that they may be remnants of preferential sites for the insertion of introns in pre-mRNA genes. Another possibility is that they might be required for efficient splicing (52). As we have demonstrated that the base-pairing regions are conserved in yeast species, and taking as reference *S.cerevisiae* snRNAs, we can hypothesise that the non-random G-5′ site-A pattern observed in nearly 62% of the *P.angusta* and *Y.lipolytica* introns may extend the base pairing to either U1 or U6 snRNAs. Most yeast introns also exhibited a conserved AAG pattern in the region preceding the 5′ site. This pattern might be important for interacting with the T-rich loop of U5 snRNA or for enhancing the binding of the first exon to PRP8 protein (53,54). A conserved nucleotide region was also identified on both sides of the branch site. The T residue preceding this motif may act by binding *trans*-acting factors (4), whereas the A residue following this sequence may base pair to U2 snRNA (55). However, the low affinity of some of the branch sequences for the U2 sequence suggests that this base-pairing interaction alone is not the initial trigger for aligning the U2 molecule with the pre-mRNA. A clear preference was also noted for either T or A immediately upstream of the 3′ site. This has also been observed in *S.pombe* (46), but not in other eukaryotes: G-YAG in plants and N-CAG in mammals (47,56). Thus, preferential splice motifs and their nucleotide environments are species dependent.

PolyT tracts are known to occur in metazoan introns (3,57), as well as in *S.pombe* (58) and *S.cerevisiae* (11,49) introns, notably in the region between the branch site and the 3′ splice site (S2). These patterns were searched (J.van Helden, unpublished results) using the program position-analysis (59). We confirmed that *S.cerevisiae* introns have an over-representation of polyT elements at strategic sites (see MIPS), i.e. not only in the S2 region as previously reported (11,49) but also in the immediate region upstream from the branch site (J.van Helden, unpublished results). Although not clearly demonstrated, such bias is strongly suspected in the *Saccharomyces–Kluyveromyces* species and in *C.tropicalis* on the basis of visual investigations (see MIPS). However, additional sequencing data are required to prove that such bias is statistically significant in the other hemiascomycetous species. This suggests that yeast and metazoan introns may share more similarities than previously assumed. The role of T-rich sequences in intron recognition has been well documented in metazoa. Such sequences are essential for boosting efficient splicing and may help to distinguish appropriate splice sites from cryptic sites (3,57).

Despite the fact that they share a similar exon/intron architecture, illustrating their common origins, each hemiascomycetous yeast species seems, however, to have developed its own intron signature to optimise splicing efficiency and possibly gene expression.

Origin of spliceosomal introns in yeasts

The high degree of conservation of the major spliceosomal components (snRNAs and many splicing factors) between hemiascomycetes and the other metakaryotes (e.g. animals, plants, fungi, protozoa) confirms that the splicing machinery has been evolutionarily conserved in this superkingdom (28,29), and suggests an ancient origin probably pre-existing the first metakaryote. There are, however, marked differences between the hemiascomycetous yeast U1, U2 and U5 snRNAs and their counterparts in *S.pombe* and metazoans. There are also differences among splicing factors, some being present in *S.pombe* and humans but absent in *S.cerevisiae*, some being present in humans but not in yeasts, and several others being probably unique to *S.cerevisiae* (58). Thus, Srp1p and Srp2p, which are two proteins belonging to the family of SR splicing factors, are present in *S.pombe* but not in *S.cerevisiae* (20). Although further investigations are required to confirm our findings, preliminary analyses (based on sequence and functional domain conservation) show that a Srp1-like protein may exist in hemiascomycetes, except in *S.cerevisiae* that probably lost it. All these observations indicate that a part of the splicing machinery has diverged in the hemiascomycetes, and that *S.pombe* probably exhibits the archetype of the splicing machinery (58).

As previously mentioned, our data are compatible with the hypothesis that most present day intron positions in hemiascomycetous yeasts, notably those occupying exactly the same position in the aligned homologues, are relics of ancestral positions (45,60–62). The analyses of intron positions among homologous yeast genes reveal that several highly conserved homologues (i.e. those with non-ambiguous alignments that do not contain any insertion/deletion in the zone of interest), exhibit different intron positions. We cannot exclude that these intron positions result from intron sliding or intron gain events (63). The intron sliding scenario was suspected when the orthologous intron positions are separated by 2 or 3 nt on the scaffold gene. A series of examples is available at the MIPS intron website. Thus, in the *S.pombe* YJL166w homologue, an intron interrupts a CTA codon between the second and third positions, whereas in *P.angusta*, the apparently orthologous intron interrupts a GGC codon, 2 nt downstream of the *S.pombe* position. Similarly with the

YGR001C case described in the results in which the suspected orthologous introns are inserted 3 nt apart and within different codons, i.e. a leucine codon (TT|G) in *S.cerevisiae* versus a glutamine codon (CA|A) in *Y.lipolytica*. Similar comparisons were performed with the YDR367w homologues where one intron appeared to be inserted between the codons encoding QK|SF in *S.pombe*, whereas its counterpart appears to be inserted between Q|KSF in *D.hansenii*. In some cases, these divergent intron positions were found located in highly divergent regions, like in the YBL099w case, making it thus difficult to know whether these discrepancies result from a simple artefact of alignment (63). No significant sequence conservation between these supposed 'orthologous' introns could be detected, however, which makes it difficult to discriminate between differential intron sliding events, intron loss or even independent intron insertion events.

Our data are compatible with the hypothesis that introns were massively lost during evolution in the hemiascomycetes, but do not exclude the possibility that intron sliding or intron gain events might have occurred since they separated from the last common ancestor. As evidenced by the increasing number of works reporting similar findings in eukaryotic organisms (61,62,64,65), it is now admitted that both intron losses and intron gains can occur during evolution.

AVAILABILITY OF RESULTS

The data are available at the MIPS website (http://mips.gsf.de/proj/yeast/reviews/intron/) as part of the Comprehensive Yeast Genome Database project (CYGD).

ACKNOWLEDGEMENTS

We are grateful to Bertrand Séraphin for reviewing the MIPS intron website, and encouraging us. We are also very grateful to Véronique Martin, Igor Tetko and Andrée Lepingle for their help with the statistical analyses and sequence annotations, respectively. This work was supported by INRA, CNRS and the GDR/CNRS 2354 'Génolevures II' project. E.B. was supported by an EEC grant (QLRI-1999-01333). Part of this work was also supported by a BRG grant (Ressources Génétiques des Microorganismes No 11-0926-99).

REFERENCES

1. Csank,C., Taylor,F.M. and Martindale,D.W. (1990) Nuclear pre-mRNA introns: analysis and comparison of intron sequences from *Tetrahymena thermophila* and other eukaryotes. *Nucleic Acids Res.*, **18**, 5133–5141.
2. Rymond,B. and Rosbash,M. (1992) In Broach,J.R., Pringle,J. and Jones,E.W. (eds), *The Molecular and Cellular Biology of the Yeast Saccharomyces cerevisiae*. Cold Spring Harbor Laboratory Press, Cold Spring Harbor, New York, Vol. 2, pp. 143–192.
3. Lorkovic,Z.J., Wieczorek Kirk,D.A., Lambermon,M.H. and Filipowicz,W. (2000) Pre-mRNA splicing in higher plants. *Trends Plant Sci.*, **5**, 160–167.
4. Staley,J.P. and Guthrie,C. (1998) Mechanical devices of the spliceosome: motors, clocks, springs, and things. *Cell*, **92**, 315–326.
5. Will,C.L. and Lührmann,R. (2001) Spliceosomal UsnRNP biogenesis, structure and function. *Curr. Opin. Cell Biol.*, **13**, 290–301.
6. Stevens,S.W., Ryan,D.E., Ge,H.Y., Moore,R.E., Young,M.K., Lee,T.D. and Abelson,J. (2002) Composition and functional characterization of the yeast spliceosomal penta-snRNP. *Mol. Cell*, **9**, 31–44.
7. Sharp,P.A. and Burge,C.B. (1997) Classification of introns: U2-type or U12-type. *Cell*, **91**, 875–879.
8. Deutsch,M. and Long,M. (1999) Intron-exon structures of eukaryotic model organisms. *Nucleic Acids Res.*, **27**, 3219–3228.
9. Vinogradov,A.E. (1999) Intron-genome size relationship on a large evolutionary scale. *J. Mol. Evol.*, **49**, 376–384.
10. Logsdon,J.M.,Jr (1998) The recent origins of spliceosomal introns revisited. *Curr. Opin. Genet. Dev.*, **8**, 637–648.
11. Spingola,M., Grate,L., Haussler,D. and Ares,M.,Jr (1999) Genome-wide bioinformatic and molecular analysis of introns in *Saccharomyces cerevisiae*. *RNA*, **5**, 221–234.
12. Lopez,P.J. and Seraphin,B. (2000) YIDB: the Yeast Intron DataBase. *Nucleic Acids Res.*, **28**, 85–86.
13. Derr,L.K., Strathern,J.N. and Garfinkel,D.J. (1991) RNA-mediated recombination in *S.cerevisiae*. *Cell*, **67**, 355–364.
14. Wood,V., Gwilliam,R., Rajandream,M.A., Lyne,M., Lyne,R., Stewart,A., Sgouros,J., Peat,N., Hayles,J., Baker,S. *et al.* (2002) The genome of *Schizosaccharomyces pombe*. *Nature*, **415**, 871–880.
15. Génolevures. (2000) Genomic exploration of the hemiascomycetous yeasts. *FEBS Lett.*, **487**, 1–149.
16. Blandin,G., Ozier-Kalogeropoulos,O., Wincker,P., Artiguenave,F. and Dujon,B. (2000) Genomic exploration of the hemiascomycetous yeasts: 16. *Candida tropicalis*. *FEBS Lett.*, **487**, 91–94.
17. Blandin,G., Llorente,B., Malpertuy,A., Wincker,P., Artiguenave,F. and Dujon,B. (2000) Genomic exploration of the hemiascomycetous yeasts: 13. *Pichia angusta*. *FEBS Lett.*, **487**, 76–81.
18. Casaregola,S., Lepingle,A., Bon,E., Neuveglise,C., Nguyen,H., Artiguenave,F., Wincker,P. and Gaillardin,C. (2000) Genomic exploration of the hemiascomycetous yeasts: 7. *Saccharomyces servazzii*. *FEBS Lett.*, **487**, 47–51.
19. Llorente,B., Malpertuy,A., Blandin,G., Artiguenave,F., Wincker,P. and Dujon,B. (2000) Genomic exploration of the hemiascomycetous yeasts: 12. *Kluyveromyces marxianus* var. *marxianus*. *FEBS Lett.*, **487**, 71–75.
20. Gross,T., Richert,K., Mierke,C., Lutzelberger,M. and Kaufer,N.F. (1998) Identification and characterization of srp1, a gene of fission yeast encoding a RNA binding domain and a RS domain typical of SR splicing factors. *Nucleic Acids Res.*, **26**, 505–511.
21. Altschul,S.F., Madden,T.L., Schaffer,A.A., Zhang,J., Zhang,Z., Miller,W. and Lipman,D.J. (1997) Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res.*, **25**, 3389–3402.
22. Pearson,W.R. and Lipman,D.J. (1988) Improved tools for biological sequence comparison. *Proc. Natl Acad. Sci. USA*, **85**, 2444–2448.
23. Higgins,D.G., Thompson,J.D. and Gibson,T.J. (1996) Using CLUSTAL for multiple sequence alignments. *Methods Enzymol.*, **266**, 383–402.
24. Thompson,J.D., Gibson,T.J., Plewniak,F., Jeanmougin,F. and Higgins,D.G. (1997) The CLUSTAL_X windows interface: flexible strategies for multiple sequence alignment aided by quality analysis tools. *Nucleic Acids Res.*, **25**, 4876–4882.
25. Saitou,N. and Nei,M. (1987) The neighbor-joining method: a new method for reconstructing phylogenetic trees. *Mol. Biol. Evol.*, **4**, 406–425.
26. Page,R.D. (1996) TreeView: an application to display phylogenetic trees on personal computers. *Comput. Appl. Biosci.*, **12**, 357–358.
27. Rodriguez-Medina,J.R. and Rymond,B.C. (1994) Prevalence and distribution of introns in non-ribosomal protein genes of yeast. *Mol. Gen. Genet.*, **243**, 532–539.
28. Guthrie,C. and Patterson,B. (1988) Spliceosomal snRNAs. *Annu. Rev. Genet.*, **22**, 387–419.
29. Roiha,H., Shuster,E.O., Brow,D.A. and Guthrie,C. (1989) Small nuclear RNAs from budding yeasts: phylogenetic comparisons reveal extensive size variation. *Gene*, **82**, 137–144.
30. Tang,Z., Kaüfer,N.F. and Lin,R.J. (2002) Interactions between two fission yeast serine/arginine-rich proteins and their modulation by phosphorylation. *Biochem. J.*, **368**, 527–534.
31. Malpertuy,A., Tekaia,F., Casaregola,S., Aigle,M., Artiguenave,F., Blandin,G., Bolotin-Fukuhara,M., Bon,E., Brottier,P., de Montigny,J. *et al.* (2000) Genomic exploration of the hemiascomycetous yeasts: 19. Ascomycetes-specific genes. *FEBS Lett.*, **487**, 113–121.
32. Ares,M.J., Grate,L. and Pauling,M.H. (1999) A handful of intron-containing genes produces the lion's share of yeast mRNA. *RNA*, **5**, 1138–1139.
33. Palmiter,R.D., Sandgren,E.P., Avarbock,M.R., Allen,D.D. and Brinster,R.L. (1991) Heterologous introns can enhance expression of transgenes in mice. *Proc. Natl Acad. Sci. USA*, **88**, 478–482.

34. Vinogradov,A.E. (2001) Intron length and codon usage. *J. Mol. Evol.*, **52**, 2–5.
35. Douglas,S., Zauner,S., Fraunholz,M., Beaton,M., Penny,S., Deng,L.T., Wu,X., Reith,M., Cavalier-Smith,T. and Maier,U.G. (2001) The highly reduced genome of an enslaved algal nucleus. *Nature*, **410**, 1091–1096.
36. Keeling,P.J., Deane,J.A. and McFadden,G.I. (1998) The phylogenetic position of alpha- and beta-tubulins from the Chlorarachnion host and Cercomonas (Cercozoa). *J. Eukaryot. Microbiol.*, **45**, 561–570.
37. Gilbert,W., Marchionni,M. and McKnight,G. (1986) On the antiquity of introns. *Cell*, **46**, 151–153.
38. Fink,G.R. (1987) Pseudogenes in yeast? *Cell*, **49**, 5–6.
39. Aravind,L., Watanabe,H., Lipman,D.J. and Koonin,E.V. (2000) Lineage-specific loss and divergence of functionally linked genes in eukaryotes. *Proc. Natl Acad. Sci. USA*, **97**, 11319–11324.
40. Klinz,F.J. and Gallwitz,D. (1985) Size and position of intervening sequences are critical for the splicing efficiency of pre-mRNA in the yeast *Saccharomyces cerevisiae*. *Nucleic Acids Res.*, **13**, 3791–3804.
41. Dhellin,O., Maestre,J. and Heidmann,T. (1997) Functional differences between the human LINE retrotransposon and retroviral reverse transcriptases for *in vivo* mRNA reverse transcription. *EMBO J.*, **16**, 6590–6602.
42. Esnault,C., Maestre,J. and Heidmann,T. (2000) Human LINE retrotransposons generate processed pseudogenes. *Nature Genet.*, **24**, 363–367.
43. Goodwin,T.J., Ormandy,J.E. and Poulter,R.T. (2001) L1-like non-LTR retrotransposons in the yeast *Candida albicans*. *Curr. Genet.*, **39**, 83–91.
44. Casaregola,S., Neuveglise,C., Bon,E. and Gaillardin,C. (2002) Ylli, a non-LTR retrotransposon L1 family in the dimorphic yeast *Yarrowia lipolytica*. *Mol. Biol. Evol.*, **19**, 664–677.
45. Baumbusch,L.O., Thorstensen,T., Krauss,V., Fischer,A., Naumann,K., Assalkhou,R., Schulz,I., Reuter,G. and Aalen,R.B. (2001) The *Arabidopsis thaliana* genome contains at least 29 active genes encoding SET domain proteins that can be assigned to four evolutionarily conserved classes. *Nucleic Acids Res.*, **29**, 4319–4333.
46. Zhang,M.Q. and Marr,T.G. (1994) Fission yeast gene structure and recognition. *Nucleic Acids Res.*, **22**, 1750–1759.
47. Brown,J.W. (1986) A catalogue of splice junction and putative branch point sequences from plant introns. *Nucleic Acids Res.*, **14**, 9549–9559.
48. Zhang,G., Taneja,K.L., Singer,R.H. and Green,M.R. (1994) Localization of pre-mRNA splicing in mammalian nuclei. *Nature*, **372**, 809–812.
49. Lopez,P.J. and Seraphin,B. (1999) Genomic-scale quantitative analysis of yeast pre-mRNA splicing: implications for splice-site recognition. *RNA*, **5**, 1135–1137.
50. Horowitz,D.S. and Krainer,A.R. (1994) Mechanisms for selecting 5′ splice sites in mammalian pre-mRNA splicing. *Trends Genet.*, **10**, 100–106.
51. Bhattacharya,D., Lutzoni,F., Reeb,V., Simon,D., Nason,J. and Fernandez,F. (2000) Widespread occurrence of spliceosomal introns in the rDNA genes of ascomycetes. *Mol. Biol. Evol.*, **17**, 1971–1984.
52. Long,M., de Souza,S.J., Rosenberg,C. and Gilbert,W. (1998) Relationship between 'proto-splice sites' and intron phases: evidence from dicodon analysis. *Proc. Natl Acad. Sci. USA*, **95**, 219–223.
53. Ares,M.,Jr and Weiser,B. (1995) Rearrangement of snRNA structure during assembly and function of the spliceosome. *Prog. Nucleic Acid Res. Mol. Biol.*, **50**, 131–159.
54. Teigelkamp,S., Newman,A.J. and Beggs,J.D. (1995) Extensive interactions of PRP8 protein with the 5′ and 3′ splice sites during splicing suggest a role in stabilization of exon alignment by U5 snRNA. *EMBO J.*, **14**, 2602–2612.
55. Parker,R., Siliciano,P.G. and Guthrie,C. (1987) Recognition of the TACTAAC box during mRNA splicing in yeast involves base pairing to the U2-like snRNA. *Cell*, **49**, 229–239.
56. Nakata,K., Kanehisa,M. and DeLisi,C. (1985) Prediction of splice junctions in mRNA sequences. *Nucleic Acids Res.*, **13**, 5327–5340.
57. Coolidge,C.J., Seely,R.J. and Patton,J.G. (1997) Functional analysis of the polypyrimidine tract in pre-mRNA splicing. *Nucleic Acids Res.*, **25**, 888–896.
58. Käufer,N.F. and Potashkin,J. (2000) Analysis of the splicing machinery in fission yeast: a comparison with budding yeast and mammals. *Nucleic Acids Res.*, **28**, 3003–3010.
59. van Helden,J., Olmo,M. and Perez-Ortin,J.E. (2000) Statistical analyses of yeast genomic downstream sequences reveals putative polyadenylation signals. *Nucleic Acids Res.*, **28**, 1000–1010.
60. Fast,N.M., Logsdon,J.M.,Jr and Doolittle,W.F. (1999) Phylogenetic analysis of the TATA box binding protein (TBP) gene from *Nosema locustae*: evidence for a microsporidia-fungi relationship and spliceosomal intron loss. *Mol. Biol. Evol.*, **16**, 1415–1419.
61. Robertson,H.M. (2000) The large srh family of chemoreceptor genes in *Caenorhabditis* nematodes reveals processes of genome evolution involving large duplications and deletions and intron gains and losses. *Genome Res.*, **10**, 192–203.
62. Boudet,N., Aubourg,S., Toffano-Nioche,C., Kreis,M. and Lecharny,A. (2001) Evolution of intron/exon structure of DEAD helicase family genes in *Arabidopsis*, *Caenorhabditis*, and *Drosophila*. *Genome Res.*, **11**, 2101–2114.
63. Stoltzfus,A., Logsdon,J.M.,Jr, Palmer,J.D. and Doolittle,W.F. (1997) Intron 'sliding' and the diversity of intron positions. *Proc. Natl Acad. Sci. USA*, **94**, 10739–10744.
64. de Souza,S.J., Long,M., Klein,R.J., Roy,S., Lin,S. and Gilbert,W. (1998) Toward a resolution of the introns early/late debate: only phase zero introns are correlated with the structure of ancient proteins. *Proc. Natl Acad. Sci. USA*, **95**, 5094–5099.
65. Venkatesh,B., Ning,Y. and Brenner,S. (1999) Late changes in spliceosomal introns define clades in vertebrate evolution. *Proc. Natl Acad. Sci. USA*, **96**, 10267–10271.