Molecular evolution of eukaryotic genomes: hemiascomycetous yeast spliceosomal introns

Elisabeth Bon, Serge Casaregola, Gaëlle Blandin1, Bertrand Llorente1, Cécile Neuveglise, Martin Munsterkotter2, Ulrich Guldener2, Hans-Werner Mewes2, Jacques Van Helden3, Bernard Dujon1 and Claude Gaillardin*

Laboratoire de Génétique Moléculaire et Cellulaire CNRS-INRA, Institut National Agronomique Paris-Grignon, F-78850 Thiverval-Grignon, France, 1Unité 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, 2MIPS, Institut für Bioinformatik, GSF-Forschungszentrum für Umwelt und Gesundheit, GmbH-, Ingolstädter Landstraße 1, D-85764 Neuherberg, Germany and 3Unité 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 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.
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 (9-11,12). 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 *Genolevures* 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 *Genolevures* RST database are listed in Table 2. The *S.cerevisiae* intronic structures like their homologues in *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 *Genolevures* project

<table>
<thead>
<tr>
<th>Yeast</th>
<th>Strains</th>
<th>Code&lt;sup&gt;a&lt;/sup&gt;</th>
<th>RST&lt;sup&gt;b&lt;/sup&gt;</th>
<th>Genomic coverage</th>
<th><em>Saccharomyces cerevisiae</em> homologues&lt;sup&gt;c&lt;/sup&gt;</th>
</tr>
</thead>
<tbody>
<tr>
<td><em>Saccharomyces servazzii</em></td>
<td>CBS 4311</td>
<td>AT</td>
<td>2570</td>
<td>0.2x</td>
<td>1410</td>
</tr>
<tr>
<td><em>Saccharomyces kluverri</em></td>
<td>CBS 3082</td>
<td>AU</td>
<td>2528</td>
<td>0.2x</td>
<td>1406</td>
</tr>
<tr>
<td><em>Kluyveromyces marxianus</em></td>
<td>CBS 712</td>
<td>AZ</td>
<td>2493</td>
<td>0.2x</td>
<td>1301</td>
</tr>
<tr>
<td><em>Pichia angusta</em></td>
<td>CBS 4732</td>
<td>BB</td>
<td>5082</td>
<td>0.4x</td>
<td>2320</td>
</tr>
<tr>
<td><em>Debaryomyces hansenii</em></td>
<td>CBS 767</td>
<td>BC</td>
<td>2830</td>
<td>0.2x</td>
<td>1119</td>
</tr>
<tr>
<td><em>Candida tropicalis</em></td>
<td>CBS 94</td>
<td>BD</td>
<td>2541</td>
<td>0.2x</td>
<td>1130</td>
</tr>
<tr>
<td><em>Yarrowia lipolytica</em></td>
<td>CLIB 89</td>
<td>AW</td>
<td>4940</td>
<td>0.4x</td>
<td>1078</td>
</tr>
</tbody>
</table>

<sup>a</sup>Code recovered at the beginning of each sequence.

<sup>b</sup>Number of RSTs.

<sup>c</sup>Number of *S.cerevisiae* homologues analysed.
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 \textit{S.cerevisiae} (or in \textit{Candida albicans} or \textit{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 ($\chi^2$) analyses. $\chi^2$ 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 \textit{S.cerevisiae}. The 5' and branching motifs absent from \textit{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 \textit{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 \textit{Saccharomyces servazzii}, 36 in \textit{Saccharomyces kluveri}, 19 in \textit{Klyveromyces marxianus}, 12 in \textit{Candida tropicalis}, 14 in...
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 $\chi^2$ 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, $\chi^2$ 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

<table>
<thead>
<tr>
<th>Spliceosomal intron types</th>
<th>S. cerevisiae</th>
<th>S. servazzii</th>
<th>S. kluyveri</th>
<th>K. marxianus</th>
<th>C. tropicalis</th>
<th>D. hansenii</th>
<th>P. angusta</th>
<th>Y. lipolytica</th>
</tr>
</thead>
<tbody>
<tr>
<td>Entire</td>
<td>260</td>
<td>141</td>
<td>140</td>
<td>130</td>
<td>130</td>
<td>113</td>
<td>113</td>
<td>119</td>
</tr>
<tr>
<td>Partial</td>
<td>-</td>
<td>14</td>
<td>10</td>
<td>6</td>
<td>5</td>
<td>11</td>
<td>11</td>
<td>11</td>
</tr>
<tr>
<td>Missing</td>
<td>-</td>
<td>10</td>
<td>6</td>
<td>5</td>
<td>13</td>
<td>17</td>
<td>17</td>
<td>12</td>
</tr>
<tr>
<td>S. pombe -like</td>
<td></td>
<td>154</td>
<td>9</td>
<td>26</td>
<td>8</td>
<td>9</td>
<td>17</td>
<td>12</td>
</tr>
<tr>
<td>New introns</td>
<td></td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>0</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Number of introns</th>
<th>260</th>
<th>141</th>
<th>140</th>
<th>130</th>
<th>130</th>
<th>113</th>
<th>113</th>
<th>119</th>
</tr>
</thead>
<tbody>
<tr>
<td>Intron location</td>
<td>RP $^a$</td>
<td>106</td>
<td>14</td>
<td>10</td>
<td>6</td>
<td>5</td>
<td>13</td>
<td>17</td>
</tr>
<tr>
<td></td>
<td>NRP $^b$</td>
<td>154</td>
<td>9</td>
<td>26</td>
<td>8</td>
<td>9</td>
<td>17</td>
<td>12</td>
</tr>
<tr>
<td>Total number of genes</td>
<td>5261</td>
<td>1410</td>
<td>1406</td>
<td>1301</td>
<td>1130</td>
<td>1119</td>
<td>2320</td>
<td>1078</td>
</tr>
<tr>
<td>Number of genes with introns</td>
<td>252</td>
<td>22</td>
<td>36</td>
<td>17</td>
<td>11</td>
<td>12</td>
<td>27</td>
<td>19</td>
</tr>
<tr>
<td>Proportion of genes with introns</td>
<td>4.8</td>
<td>1.6</td>
<td>2.6</td>
<td>1.3</td>
<td>1.0</td>
<td>1.1</td>
<td>1.2</td>
<td>1.7</td>
</tr>
</tbody>
</table>

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.

---

Table 4. CYGD intron inventory

<table>
<thead>
<tr>
<th>Spliceosomal intron types</th>
<th>S. cerevisiae</th>
<th>S. servazzii</th>
<th>S. kluyveri</th>
<th>K. marxianus</th>
<th>C. tropicalis</th>
<th>D. hansenii</th>
<th>P. angusta</th>
<th>Y. lipolytica</th>
</tr>
</thead>
<tbody>
<tr>
<td>Entire</td>
<td>260</td>
<td>141</td>
<td>140</td>
<td>130</td>
<td>130</td>
<td>113</td>
<td>113</td>
<td>119</td>
</tr>
<tr>
<td>Partial</td>
<td>-</td>
<td>14</td>
<td>10</td>
<td>6</td>
<td>5</td>
<td>11</td>
<td>11</td>
<td>11</td>
</tr>
<tr>
<td>Missing</td>
<td>-</td>
<td>10</td>
<td>6</td>
<td>5</td>
<td>13</td>
<td>17</td>
<td>17</td>
<td>12</td>
</tr>
<tr>
<td>S. pombe -like</td>
<td></td>
<td>154</td>
<td>9</td>
<td>26</td>
<td>8</td>
<td>9</td>
<td>17</td>
<td>12</td>
</tr>
<tr>
<td>New introns</td>
<td></td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>0</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Number of introns</th>
<th>260</th>
<th>141</th>
<th>140</th>
<th>130</th>
<th>130</th>
<th>113</th>
<th>113</th>
<th>119</th>
</tr>
</thead>
<tbody>
<tr>
<td>Intron location</td>
<td>RP $^a$</td>
<td>106</td>
<td>14</td>
<td>10</td>
<td>6</td>
<td>5</td>
<td>13</td>
<td>17</td>
</tr>
<tr>
<td></td>
<td>NRP $^b$</td>
<td>154</td>
<td>9</td>
<td>26</td>
<td>8</td>
<td>9</td>
<td>17</td>
<td>12</td>
</tr>
<tr>
<td>Total number of genes</td>
<td>5261</td>
<td>1410</td>
<td>1406</td>
<td>1301</td>
<td>1130</td>
<td>1119</td>
<td>2320</td>
<td>1078</td>
</tr>
<tr>
<td>Number of genes with introns</td>
<td>252</td>
<td>22</td>
<td>36</td>
<td>17</td>
<td>11</td>
<td>12</td>
<td>27</td>
<td>19</td>
</tr>
<tr>
<td>Proportion of genes with introns</td>
<td>4.8</td>
<td>1.6</td>
<td>2.6</td>
<td>1.3</td>
<td>1.0</td>
<td>1.1</td>
<td>1.2</td>
<td>1.7</td>
</tr>
</tbody>
</table>

A complete intron list is available at the MIPS website (http://mips.gsf.de/proj/yeast/reviews/intron/intron_RST.html).
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.

![Compilation of splice motifs](image)

**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. The level of significance of the χ² 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 indicated in bold, the confirmed novel motifs (encountered several times) are underlined, and putative novel splice motifs (encountered once only) are in italics.
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: AAGINNN for *S.cerevisiae*, A(\(\overline{\text{A}}/\overline{T}\))GITTA for *S.servazzii*, AANINNN for *S.kluyveri*, A(\(\overline{\text{A}}/\overline{\text{C}}\))GIGCT for *K.marxianus*, A(\(\overline{\text{A}}/\overline{\text{C}}\))GITCT for *C.tropicalis*, (\(\overline{\text{A}}/\overline{T}\))A(\(\overline{\text{A}}/\overline{T}\))\(\overline{\text{N}}\)AA for *D.hansenii*, ANGIGCT for *P.angusta* and NN(\(\overline{\text{A}}/\overline{\text{G}}\))\(\overline{\text{N}}\)AA 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 (\(\overline{\text{A}}/\overline{T}\))A(\(\overline{\text{A}}/\overline{T}\)) 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–Kluveromyces 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–Kluveromyces 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 underestimated 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
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

---

**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.
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 ($\Delta = 15$ nt). i2 and i’2 were in a different phase (GAG \| GAA for i2 and GAG \| GAA for i’2) and at a different location ($\Delta = 65$ 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 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 ($\sigma$) 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 $P = 0.05$.

![Diagram](image_url)
(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 (ATT for i1 and AITG 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
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 (PuAT4±6G-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), *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,*
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, K.lactis, K.marxianus, 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 question-able 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 excep-tional (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 *Calicibac* (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 AU 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 (52). 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 (TTTG) in S.cerevisiae versus a glutamine codon (CAAG) in Y.jlipolytica. Similar comparisons were performed with the YDR367w homologues where one intron appeared to be inserted between the codons encoding QKSF in S.pombe, whereas its counterpart appears to be inserted between QKSF 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


