Supplementary Materials

Functional analysis of 5'ss positions

Given the low conservation levels in the exonic part of the 5'ss in fungi, we sought to determine whether the exonic part of the 5'ss in fungi is of any importance in the context of binding of U1 snRNA and other factors. Such analyses have been performed in the past only for several metazoans and for *A. thaliana* (Burge and Karlin, 1997; Carmel et al., 2004; Lim and Burge, 2001; Thanaraj and Robinson, 2000), by showing the existence of a "see-saw" effect, or anti-correlation, between pairs of positions from opposing sides of the exon/intron boundary. This effect has been attributed to U1 snRNA binding: Poor matching between nucleotides at one side of the exon/intron boundary is compensated by stronger matching on the other side.

Thus, we decided to assess whether a "see-saw" effect can be observed across the 5'ss of fungi. To calculate dependency between positions of the 5'ss (positions -4 to 8), the consensus nucleotide at each of the 5'ss positions – defined as the most frequent nucleotide per position – was first determined for each organism. For every intron, a score of 1 or 0 was assigned to each position of the 5'ss, according to the position's conformity to the 5'ss consensus. This generated an N-by-P matrix, where N is the number of introns and P the number of positions in the 5'ss (in this case 12). A symmetrical P-by-P matrix containing the Spearman rank correlation coefficients between each pair of positions in the 5'ss was calculated, as well as a corresponding P-by-P matrix indicating the significance of this correlation. Each statistically significant correlation ($p<0.01$) was assigned a color-code according to its nature (positive/negative) and strength.

We found (Supplementary Figure S3) that the pattern of correlations between positions in fungi is similar to that in metazoans. A clear "see-saw" effect was observed in fungi, with positions -2 and -1 tending to anti-correlate with positions 2 to 6. In fewer cases, position -3 was found to anti-correlate with the positions as well. This indicates that despite the relatively low conservation found at these positions in fungi, these positions are nonetheless important for 5'ss selection. Notably, position -4 was usually not involved in any correlations, while positions 7 and 8 were, indicating that the latter positions tend to be involved in splice site selection, while the former is of less importance in this context.

We also observed that adjacent positions at either edge of the 5'ss (positions along the central diagonal) tended to correlate positively. Such observations are presumably due to "stacking effect" (Carmel et al., 2004), or the need to ensure good U1/5'ss pairing at the edges. Among many organisms, such positive correlations were found not only between immediately adjacent positions, but also between close positions in general. In *N. crassa* and *C. elegans* for example, such correlations can be found between pairs of positions from position 4 to position 8.

Full discussion of correlations between the 5'ss and U1-snRNA

In this section we provide a more thorough analysis and discussion of the changes found in various positions of the 5'ss, and their correlation with changes in U1 snRNA.

Variations in the 5'ss consensus were found in three pairs of positions: positions 3 & 4, positions 7 & 8 and positions -3 & -4 (presented in Figure 1). We were interested in finding to what extent this variation correlates with complementary variation in U1 snRNA (presented in Figure 2B and 2C and in Supplementary Figure 1).

***Positions 3 & 4*:** The consensus 5'ss signal for most organisms is GTAAGT. However, among most hemiascomycetous yeasts we find a 'T' at position 4 (GTATGT), while in *Y. lipolytica* we find a 'G' at position 3 (GTGAGT). When looking at the complementary position in U1 snRNA, we find that these two positions remain conserved throughout evolution, and in all cases remain 3'-CAUUCA-5'. Thus, changes in U1 snRNA cannot explain the differential consensus nucleotides at positions 3 and 4. Notably, the two variations observed in the 5'ss of hemiascomycetous fungi do not appear to be at total 'disregard' of U1 snRNA: both the 'T', at position 4, and the 'G', at position 3, can form non-Watson-Crick baseparing with the 'T' across them (Ast, 2004).

A possible candidate explaining the tendency for 'T' at position +4 in hemiascomycetous fungi is U6-snRNA. U6-snRNA has a highly conserved 'ACAGAG' sequence, the first three bases of which undergo base-pairing with the 'TGT' consensus at positions +4 to +6 in *S. cerevisiae* (Kandels-Lewis and Seraphin, 1993; Konarska et al., 2006). Thus, the balance of forces between U1 and U6 snRNA may determine the consensus nucleotide at position +4: among hemiascomycetous fungi, U6 snRNA is more dominant, and hence 'T' is the consensus nucleotide, whereas among other organisms U1 snRNA gains the upper hand, as indicated by the 'A' consensus nucleotide. However, the consensus nucleotides at position +3 and +4 in *Y. lipolytica* suggest that a more complex mechanism is at work here. In this organism, the nucleotides at both these positions are extremely conserved, but in a manner which is neither complementary to U1 snRNA (Figure 1), nor to U6 snRNA, which was identified for this organism as well and found unchanged in the sequence binding the 5'ss (data not shown).

A further partial explanation for the divergence from the nucleotides complementary to U1 snRNA, at these two positions in hemiascomycetous fungi could be a tendency to avoid hyperstabilization of the binding between U1 snRNA and the 5'ss. Staley et. al have demonstrated that hyperstabilization of the U1 snRNA/5′ss interaction by extending base pairing between U1 and the 5′ splice site can lead to temperature-sensitive splicing repression in yeast (Staley and Guthrie, 1999). Since the remaining positions are all highly conserved among hemiascomycetous fungi, the non-complementary nucleotide at position 3 or 4 may be a mechanism that avoids hyperstabilization of the base-paring between U1 snRNA and the 5'ss, and allows unwinding of U1 snRNA from the 5'ss before the first step of splicing. An alternative explanation is that it reflects binding of a different factor: Du et. al have shown that the U1-snRNP protein U1C preferentially binds a sequence of 'GTAT', and is heavily implicated in the recognition of the 5'ss. Hence, the preference for 'T' at position 4 may reflect a purifying selection on behalf of U1C (Du and Rosbash, 2002).

***Positions 7&8*:** Among the majority of organisms, the consensus nucleotides at these positions is either 'AT' or 'TT'. These positions were found to correlate with the corresponding positions in U1-snRNA in particular when the consensus nucleotide was dominant. For example, examining position +7 in the 5'ss of *C. elegans*, we noted a clear dominance of 'T' (appearing in 50% of the introns), in contrast to all other metazoans in which there is a slight preference for 'A's or for 'G's at this position. This reflects the U1-snRNA of *C. elegans*, which contains 'A' at the position base-pairing with U1-snRNA (basepairing with 'T'), in contrast to all other metazoan species in which there is a 'T' at this position (forming Watson-Crick basepairs with 'A', and non-Watson-Crick basepairs with 'G'). Similarly, in several yeasts and protozoans the consensus nucleotide at position +7 (either T or A) is extremely conserved. Such is the case in in *Y. lipolytica* (the consensus nucleotide is 'A', appearing in 75% of the introns), *D. hansenii* (consensus 'A' in 61% of the introns) and in *C. parvum* (consensus 'T' in 84% of the introns). In all these cases, this correlates with the corresponding position in U1-snRNA. In position +8, the consensus nucleotide of most organisms is 'T', again correlating with the 'A' in the corresponding position in U1-snRNA. However, in other organisms the consensus appears to be determined by factors other than U1-snRNA. For example, in *C. glabrata* and *S. cerevisiae* the clear consensus nucleotides at position +7 is 'T' (appearing in 58% and 44% of the introns, respectively). However, based on the corresponding position in U1-snRNA we would expect this consensus to be an 'A'.

***Positions -3 & -4*:** In the functional analysis of the 5'ss positions, we found that among most organisms these two positions were not involved in anti-correlations with intronic positions of the 5'ss, indicating that they are not important in the context of U1 binding. However, in six organisms these positions were involved in such anti-correlations: in the three mammals (dog, mouse, human), *S. pombe*, *C. neoformans* and *D. discoideum*. In all these organisms with the exception of *D. discoideum*, the consensus nucleotide was found to anti-correlate with the coresonding nucleotide in U1-snRNA. Thus, in the three mammals the consensus nucleotide is 'C', corresponding to the 'G' in their U1-snRNAs, whereas in *S. pombe* and *C. neoformans* the consensus nucleotide is 'A', corresponding to the 'T' in U1-snRNA. This conclusion is strengthened by the fact that in various organisms in which no such complementarity exists (such as the four non-mammalian metazoans, and *A. thaliana*), position -3 was not found to participate in anti-correlations.

In position -4, the consensus nucleotide is usually an 'A'. This does not correlate with the corresponding nucleotide in U1-snRNA, and is in line with our results for the functional analysis, in which we found that position -4 is usually not involved in anti-correlations with intronic positions. These findings suggest that further factors may be of importance in this context. One such factor might be U5-snRNA: indeed, studies in the past have noted the conservation of A's at position -2 to -4 among some hemiascomycetous fungi and suggested that a complementary stretch of 'TTT' in U5-snRNA is capable of binding it (Long et al., 1997; Lopez and Seraphin, 1999; Newman and Norman, 1992; Spingola et al., 1999). Thus, the preference for 'AAA' at position -2 to -4, among most organisms, may reflect relative dominance of U5-snRNA, while the change from 'A' to 'C' at position -3 may reflect an increase in the dominance of U1 snRNA.

We concluded that while the core of the 5'ss consensus, between position -1 and position +6, is, indeed, determined by U1 snRNA binding, the final 5'ss consensus is determined by the integrated preferences of U1 snRNA along with several further factors, possibly including U1C, U5 snRNA and U6 snRNA.

Correlation between intron length and PPT strength

Examining the PPT enrichment indexes among the different organisms, we noted that PPTs tend to appear in organisms characterized by longer introns. Moreover, this tendency is not confined to specific phylogenetic groups, but can be observed across all organisms. Indeed, as can be seen in Supplementary Figure S5, a very high correlation exists between the PPT enrichment index and median intron length (Pearson correlation, r=0.9, p=2.85e-07). Moreover, in 11 of the 22 organisms positive, albeit weak, correlations were also found between the PPT enrichment indexes and intron lengths within the same genomes (data not shown), altogether suggesting an increased role for the PPT in longer introns.

This prompted us to examine whether the increased PPT enrichment index found among vertebrates, relative to fungi, can be attributed to the fact that these introns are generally longer and can therefore harbor longer PPTs, in contrast to the shorter fungal introns. We addressed this question by compiling datasets of short introns (<200 nt) of each of the five vertebrates. However, despite the fact that the median intron length in these datasets was ~100 nt, on a similar scale as the median intron lengths found among the various fungi, the PPT enrichment index in these introns remained considerably higher than the one observed among fungi (data not shown). This demonstrates that the bias for PPTs is not merely as a result of an intron length bias, but presumably in accordance with a biological requirement for the presence of a PPT.

Taken together, these results suggest that PPTs are of increased importance for recognizing long introns. However, once an organism is already generally characterized by long introns with PPTs, the PPT also becomes essential in its shorter introns, presumably because the organism's splicing machinery has become adapted to rely on it.

Analysis of nucleotide composition of the PPT

The algorithm for identifying PPTs did not distinguish between the two pyrimidines ('T' and 'C') nor between the two purines ('A' and 'G'). Therefore, comparing the absolute ratio between the occurrences of these pairs of nucleotides within PPTs is informative since it can indicate whether there is selection for, or against, certain nucleotides. We examined these two ratios among all organisms. The absolute T:C ratio was greater than 1 among all organisms in which statistically significant PPTs were found, indicating a bias for T. These results are consistent with (Bouck et al., 1995; Coolidge et al., 1997).

As the introns of most organisms are more rich in 'T' than in 'C' to begin with, we adjusted the absolute ratio to the background T:C ratio within the intron, by dividing the former by the latter, yielding the adjusted T:C ratio. Among all non-metazoans, the adjusted T:C ratio still showed a bias for 'T', usually in a highly significant manner (χ^2, $p<0.01$). However, among metazoans we observed a gradual decrease in this ratio: the

adjusted 'T':C ratio decreases from 1.5:1 to 1.18:1, 1.16:1, 1.07:1 and 0.94:1 in *C. elegans*, *D. melanogaster*, zebrafish, chicken and human respectively. Thus, in dog, mouse and human there is even a slight but significant over-representation of C vs. T, with respect to the ratio between these two nucleotides throughout the introns. This result is consistent with our observation in the bias-plot analysis, showing an increase in 'C' throughout the metazoan lineages. See Supplementary Table S2 for full results on the above analyses.

We next performed a similar analysis, examining the G:A ratio, in order to assess whether there was a selection against As, in the PPTs of vertebrates, as we had noted in the bias plot analysis. We found both the absolute, and adjusted G:A ratio to favor 'G's in a highly significant manner (χ^2, p~0). Among other organisms, this selection against 'A', or for 'G', was not consistently detected (see Supplementary Table S2).

In the bias plot analysis of all metazoans excluding *C. elegans*, position -10 was found to serve as a key position, with a relative 'C' bias appearing downstream of it, and a 'T' bias reaching its peak at this position. In order to verify that nucleotide composition does, indeed, differ upstream and downstream of this position, each PPT was divided into two segments: The upstream segment, including all positions in the PPT upstream of, and including, position -10, and the downstream segment, including all positions downstream of position -10. For each organism, the overall nucleotide composition of the upstream segments was compared to that of the downstream segments. This analysis was applied to the six metazoans, from *D. melanogaster* to human, and confirmed the following:

1. The downstream segment of the PPT was found to contain a higher 'C' content than the upstream segment. In all six organisms, the 'C' content in the upstream segment was 6%-9% lower than in the downstream one, and was found to be highly significant (χ^2, p≈0). In human, for example, the 'C' content of the upstream segment was 35.7%, while that of the downstream one was 42.6%.
2. The frequency of 'A' in the upstream segment was more than double that of the upstream segment, indicating a bias against 'A' near the 3'ss. In human, for example, the A-content of the upstream segment was 4.8%, while that of the downstream segment was 2.3%. Here, too, the differences were found to be highly significant in all cases (χ^2, p≈0 for all organisms).

These results confirmed the non-homogenous nucleotide composition of the PPT, and supported our observations of two different signals being located across it.

Factors binding the PPT

Here we provide a detailed explanation of the analysis pertaining to the factors binding the PPT.

We set out to determine to what extent changes in the PPT are determined by corresponding changes in the splicing factors that bind the 3' end of introns during early stages of splicing. Specifically, we focused on U2AF65 and U2AF35, which recognize the PPT and the 3'ss, respectively (Kent et al., 2005; Zamore and Green, 1989), and on SF1, which binds the BS and facilitates the binding of U2AF65 to the adjacent PPT (Manceau et al., 2006). We concentrated on the functional residues in these proteins.

These include the regions that are important for RNA binding, as well as residues that are important for interactions with other splicing factors. Specifically, U2AF65 comprises an arginine-serine-rich region (RS-domain) at the N-terminal; two RNA recognition domains (RRM), RRM1 and RRM2 that bind the PPT; and a third RRM domain at the C-terminal called the U2AF homology motif (UHM) (Kielkopf et al., 2004). These RRM domains contain two motifs, ribonucleoprotein 1 and ribonucleoprotein 2 (RNP1 and RNP2, respectively), which are essential for their function (Maris et al., 2005). U2AF35 has a UHM at the N-terminal and an RS domain at the C-terminal. Flanking the UHM there are two zinc-finger motifs, which are crucial for its function (Webb and Wise, 2004). The interaction of U2AF35 with U2AF65 is mediated by the interaction of a tryptophan located at the N-terminal of U2AF65 with a hydrophobic pocket within the UHM domain of U2AF35. SF1 has a K-Homology (KH) domain, which binds the BS; a motif consisting of two adjacent Serine-Proline residues, termed SPSP motif, which can undergo phosphorylation and thereby enhances the interaction with U2AF65; and a tryptophan, that is located near the SPSP motif and interacts with the hydrophobic pocket of the UHM of U2AF65, thereby mediating the physical interaction between SF1 and U2AF65 (Manceau et al., 2006; Selenko et al., 2003). In our analysis we focused on all the above-described functional regions and residues.

Using sequence searches and protein domain analysis (see Compilation of U2AF65, U2AF35 and SF1 datasets) we searched for homologs of these three proteins in all 22 organisms. For all the metazoans, plants, non-hemiascomycetous fungi, and *D. discoideum* we found homologs of all three proteins (see Supplementary tables S7, S8 and S9). In U2AF65, the three RRM domains (RRM1, RRM2, and UHM) were present in all homologs, and their functional residues were all found to be conserved as well. Such was also the case for the RNP1 and RNP2 motifs of RRM1and RRM2 (see Supplementary Figures S7 and S8), the hydrophobic pocket in the UHM domain of U2AF65 (see Supplementary Figures S9), and the tryptophan at the N-terminal region (see Supplementary Figure S10). In U2AF35, we found a conserved hydrophobic pocket and zinc-finger motifs (see Supplementary Figure S11). Finally, in SF1 we found the KH-domain and SPSP motif to be conserved (see Supplementary Figure S12). We concluded that in the analyzed metazoans, plants, non-hemiascomycetous fungi, and in *D. discoideum*, recognition of the 3'ss is likely to take place as in human, with the U2AF heterodimer interacting with SF1 in the recognition of the 3' intron end. These results are in line with our findings for the PPT, because in all these organisms statistically significant PPT enrichment indexes were found.

We next analyzed the RRMs, binding the PPT, in greater detail. Comparing the RRMs of U2AF65 among species to the corresponding human RRM (Figure 5A), we observed two phenomena. First, the RRM2 domain is more conserved in vertebrates and fungi, with respect to human, than RRM1 and UHM. These results suggest that RRM2 may be the dominant domain in terms of PPT binding, among non-metazoans (see Discussion). Second, we observed that among vertebrates there is almost 100% identity conservation in RRM1 and RRM2 with respect to human. This conservation gradually decreases from vertebrates to invertebrates, and even more among fungi. This decreasing gradient correlates with the trend observed in the PPT, which was found to be weaker in

invertebrates than in vertebrates, and even weaker among most fungi. While these results may suggest that the PPT coevolved with RRMs binding it, the decreased conservation may also reflect increased phylogenetic distances.

Thus, to assess the functional importance of the decreased conservation, we decided to focus on specific, key residues on RRM1 and RRM2 that have previously been shown to be required for PPT binding in human (Sickmier et al., 2006). These included residues participating in main-chain, side-chain, and water-mediated interactions (Sickmier et al., 2006). The characteristics of these residues, in terms of polarity, charge, and aromaticity are therefore important for the ability of U2AF65 to bind the PPT. A change in polarity will presumably affect the water-mediated interactions, whereas any change in charge, polarity or aromaticity is expected to affect the side-chain interactions. Among non-hemiascomycetous fungi, we identified many such changes, with respect to metazoans, were found, in key residues both in RRM1 (Figure 5B) and in RRM2 (Figure 5C). These results suggest that the decrease in PPT strength among fungi, relative to metazoans, is linked to detrimental changes in key residues on U2AF65 required for PPT binding. This conclusion is strengthened by the fact that relatively much fewer changes were observed in the RRMs of *D. discoideum* and *A. thaliana*, despite the fact that phylogenetically they are more distant from metazoans than non-hemiascomycetous fungi. This correlates with our findings pertaining to the PPT, which is stronger, in these two organisms, than among non-hemiascomycetous fungi (see Discussion).

The hemiascomycetous fungi present a more divergent pattern in their examined splicing factors. They can be separated into two groups: *S. Cerevisiae*-like, and non-*S. cerevisiae* like. In *S. cerevisiae*, MUD2, an analog of U2AF65 is part of the commitment complex and contacts the pre-mRNA during the commitment complex assembly (Abovich et al., 1994). MUD2 has only one RNA binding domain, and interacts directly with MSL5 (Rutz and Seraphin, 1999), the SF1 analog in *S. cerevisiae* that recognizes the BS, and with U2 snRNP (Abovich et al., 1994) during splicing. The hemiascomycetous fungi *C. glabrata*, *E. gossypii* and *K. lactis* are cerevisiae-like: They all contain a single copy of a MUD2 homolog (see Supplementary Figure S13 and Supplementary Table S8). These organisms have homologs of MSL5 as well, with a conserved KH-domain and SPSP motif (see Supplementary Figure S14). Moreover, no functional homologs of U2AF35 were found among these organisms, as in *S. cerevisiae*. Notably, in *E. gossypi* a U2AF35 homolog was found, but it lacked the essential zinc fingers. Moreover, its open reading frame was disrupted by a stop codon, suggesting that it is a pseudogene. We concluded that for this subgroup of species, the recognition of the BS and PPT presumably takes place as is known for *S. cerevisiae*, mediated by a MUD2 homolog but not by U2AF35.

On the other hand, in *Y. lipolytica* and *D. hansenii*, two other members of the hemiascomycetous fungi, we did not find any MUD2 homologs, but found U2AF65 homologs instead. However, these homologs present several critical differences with respect to U2AF65: *D. hansenii* completely lacks the RRM1 and RRM2 domains, while in *Y. lipolytica* the essential RNP1 and RNP2 motifs are not conserved (see Supplementary Figure S15). Thus, in these two species, the U2AF65 homolog lacks the capability to bind to the PPT. Interestingly, both species have homologs of U2AF35 and

SF1, both of which appear to have retained their functionality based on the conserved UHM and zinc-finger domains in U2AF35 (see Supplementary Figure S11), and the conserved KH-domains and SPSP motif in SF1 (see Supplementary Figure S12). Moreover, both proteins have retained the ability to interact with U2AF65: The hydrophobic pocket of U2AF35 is conserved as well as the tryptophan in SF1 (see Supplementary Figures S11 and S12). However, the hydrophobic pocket of the UHM domain in the U2AF65 homolog is mutated (see Supplementary Figure S16). We concluded that in *Y. lipolytica* and *D. hansenii* the recognition of the 3'ss and BS is likely to be performed by U2AF35 and SF1, respectively, and that U2AF65 may function as a bridge between both proteins. These results agree with our findings pertaining to the PPT analysis, as in both organisms we found no PPT between the BS and the 3'ss.

Finally, *C. parvum* presents a puzzling case. In this organism we found two U2AF65 homologs, both of which have a UHM domain but no arginine-rich region at the N-terminal. One of the homologs has a further conserved RRM, but no tryptophan for the interaction with U2AF35 and a very degenerate hydrophobic pocket in the UHM. The other homolog has a tryptophan for the interaction with U2AF35, but a slightly mutated hydrophobic pocket. The U2AF35 and SF1 homologs were not fully conserved in terms of functional residues as well: The U2AF35 has zinc-finger domains, but its UHM domain lacks the hydrophobic pocket for the interaction with U2AF65 (see Supplementary Figure S11). The SF1 homolog has a KH-domain, as well as the tryptophan relevant for the interaction with U2AF65, but an EPSP motif instead of SPSP (see Supplementary Figure S12). Taken together, these results suggest that in this organism, U2AF35 may not function jointly with U2AF65.

Validation of the BS

In order to validate the results obtained by our algorithm for detecting the BS, we implemented two further algorithms that have been used in the past for detecting BSs. We found a large degree of congruence between the BSs extracted by our algorithm and the ones extracted by two other previously published BS detection methods. The congruence observed was both in terms of the identified BS motifs and the distribution of the BS distance from the 3'ss. Specifically, we have implemented the algorithms of (Kupfer et al., 2004), which was used to identify branch sites in five fungi, and the algorithm of (Kol et al., 2005), which was designed for BS detection in human and mouse. The high congruence among the three methods suggests that the results obtained are not very sensitive to the BS detection method.

Further validation of our results is obtained from the distribution of the distances between the BS and the 3'ss, in each organism. Since our algorithm gives preference to BSs located close to the 3'ss, when it is applied to a random dataset the histogram is positively skewed, peaking at the last position. However, among all organisms (excluding *Y. lipolytica*, which is discussed in detail in Results and in Discussion), the peak of the BS distribution is not immediately upstream of the 3'ss, but situated a variable number of positions upstream of the 3'ss (see Supplementary Figure S6), in line with the expectations regarding the BS.

To what extent is this algorithm, based on fungal BSs, applicable to metazoan introns? To asses this, we used a dataset of 19 introns containing biologically proven BSs, compiled by (Kol et al., 2005). Of the 16 putative BS identified by our program (3 were discarded), 9 corresponded to the biologically proven BS. Based on these results, 56.2% of the BS predictions of our algorithm, in metazoans, are exact. Closer examination of the dataset of Kol et. al reveals that of the 19 introns, 14 originate from mammals, while 5 of them have been introduced into mammalian genomes by viruses. The 5 viral BSs are atypical, and our algorithm generally failed at correctly identifying these BSs. Discarding these 5 viral introns and leaving only the introns originating from mammals, the exactitude of the algorithm increases to 72.7%. As BSs, introduced by viruses, might have unique characteristics, we estimate that between 56.2%-72.7% of the putative BSs identified by our algorithm for metazoans correspond to the biologically validated BS.

Analysis of the 3' splice site

In this analysis, we examined the last 4 positions within the intron (positions -4 to -1) and the first 2 positions in the downstream exon (positions 1 and 2). Sequence motifs of these positions are presented in Supplementary Figure S4. In position -3 we found a clear preference for either 'T' or 'C', with some organisms showing a clear preference for one nucleotide and others for the other. The preference for pyrimidines at this position has been noted before, in different organisms (Abril et al., 2005; Black, 2003; Dou et al., 2006; Smith et al., 1993). In addition, at position -3 a particularly strong selection against 'G' was observed: This nucleotide is invariably the least frequent nucleotide at this position. These finding are in line with previous studies that have found that 'G' at position -3 is particularly detrimental for splicing (Lev-Maor et al., 2003). A preference for 'G' and 'T' was observed in position 1 and 2 of vertebrate introns, respectively, consistent with previous reports (e.g. (Abril et al., 2005; Lim and Burge, 2001)). Among other organisms, the preferences were more variable but tended to include 'T' at position 2.

Compilation of U2AF65, U2AF35, and SF1 datasets

We downloaded the genomic and proteomic sequences of *Homo sapiens* (NCBI36), *Mus musculus* (NCBIM36), *Canis familiaris* (BROADD2), *Gallus gallus* (WASHUC2), *Danio rerio* (ZFISH6), *Xenopus tropicalis* (JGI4.1), *Caernohabditis elegans* (WB170), and *Saccharomyces cerevisiae* (SGD1.01) from the EnsEMBL website (http://www.ensembl.org/). The information about fungal species was obtained from the Resources for Fungal Comparative Genomics (http://fungal.genome.duke.edu/). The genomic and transcriptomic sequences for *Kluyveromyces lactis* (Klla-GL2r2), *Candida glabrata* (Cagl-GL2r2), *Debaryomyces hansenii* (Deha-GL2r2) and *Yarrowia lipolytica* (Yali-GL2r2) were downloaded from the *Génolevures* project website (http://cbi.labri.fr/Genolevures/). The information for *Eremothecium gossypii* (AGD3.0) was collected from the Ashbya Genome Database website (http://agd.vital-it.ch/). The genomic and proteomic sequences for *Neurospora crassa* (BROAD3 assembly 7), *Magnaporthe grisea* (Assembly release 5.0), *Aspergillus nidulans* (Assembly release 4), *Ustilago maydis* (Assembly release 2) and *Rhizopus oryzae* (Assembly release 3) were downloaded from the Broad Institute website. Data for *Cryptococcus neoformans* JEC21 (TIGR) was downloaded from the TIGR database (http://www.tigr.org/), and the data for

Aspergillus fumigatus (GeneDB) and *Schizosaccharomyces pombe* (SANGER1) was downloaded from the Sanger Institute website (http://www.sanger.ac.uk/projects/). Finally, data for the two protozoans *Cryptosporidium parvum* (Build 1.1) and *Dictyostelium discoideum* (Build 2.1) was downloaded from the National Center for Biotechnology Information website (http://www.ncbi.nlm.nih.gov/).

To identify the genes and protein sequences of the relevant splicing factors in the different organisms, we first extracted the known sequences of U2AF65, U2AF35, and SF1 in human; of MUD2 and MSL5 in *S. cerevisiae*; and of U2AF59 (PRP2), U2AF35, and SF1 (BPB1) in *S. pombe*. These sequences were used as queries to search for matches in the available proteomic sequences using BLASTP (Altschul et al., 1990) and Exonerate (Slater and Birney, 2005). If no matches were found in the proteomic data, we used the same query to identify the proteins in the genomic sequence using TBLASTN, Exonerate, and GeneWise (Birney et al., 2004). For all the positive matches we used Pfam (http://www.sanger.ac.uk/Software/Pfam/) or PROSITE (http://ca.expasy.org/prosite/) to confirm the existence of the characteristic domains in the three proteins. Subsequently, we performed multiple alignments of the entire protein sequences and of the domains independently, using T-COFFEE (Notredame et al., 2000), and verified the conservation of the relevant amino acid motifs. Candidates were also reciprocally compared with the human and yeast proteomes to confirm that we had identified the correct orthologous sequence. In the search for U2AF65, we did not find any candidates using the proteomic or the genomic approach for *K. lactis*, *C. glabrata,* and *E. gossypii*, but we were able to find homologs for MUD2. In the U2AF35 analysis of *E. gossypii* we found a match that lacks the first zinc-finger region in the N-terminal. We checked the upstream region of the nucleotide sequence of the gene, using GeneWise to align it against the U2AF35 from *S. pombe.* We observed that the sequence for the zinc-finger is present, but disrupted by a stop codon. On the other hand, we verified that the upstream region could potentially contain an intron, but in this case the zinc-finger will not be functional. We concluded that the U2AF35 homolog in *E. gossypii* is probably a pseudogene. Finally, for *G. gallus* we did not find matches for U2AF65, but found a GeneIndex entry in http://compbio.dfci.harvard.edu/ (TC284781) that corresponds to the incomplete mRNA for U2AF65. This cDNA has no alignment to the current chicken genome assembly. On the other hand, an alignment of the Xenopus U2AF65 protein to this cDNA showed 100% sequence conservation at the protein level. We therefore used the Xenopus protein for the subsequent protein analyses. Using the multiple alignments from each RRM type in U2AF65 we calculated the percentage of identity of each of the domains in each species compared to its human counterpart. These values are shown in Figure 5A.

Compilation of U2 snRNA dataset

We first tried to detect the U2 snRNAs genes in the different genomes using the known U2 snRNAs from human, *S. cerevisiae*, *S. pombe*, *C. neoformans* and *D. melanogaster* as queries in BLASTN searches against the entire genomes of all the other organisms. The 3 highest scoring hits were extracted, and run through the Infernal package, which was downloaded from the Rfam website (http://www.sanger.ac.uk/Software/Rfam/ftp.shtml) (Griffiths-Jones et al., 2005). Finally, we selected the best matches yielded by the

Infernal package and extracted the sequences spanning the alignment with the known U2 snRNA sequences. Altogether we were able to identify the U2-snRNA sequences in all 22 organisms.

Compilation of U1 snRNA dataset

We first tried to detect U1 snRNAs using the same approach used to find the U2 snRNAs. We used the known U1 snRNAs from human, *S. cerevisiae*, *S. pombe*, *D. melanogaster*, *A. thaliana* and *C. elegans* as queries in BLASTN searches against the entire genomes of the other organisms. However, as various regions in this molecule have undergone considerable changes throughout evolution (Roiha et al., 1989), we were unable to find relevant matches for most cases. In fact, we could only identify matches in *C. glabrata*, *K. lactis* and *E. gossypii*. These three species, like *S. cerevisiae*, have a U1 sequence longer than the metzoan U1s, and with a different secondary structure (Kretzner et al., 1990).

Next, we tried a second approach consisting of a PERL program that searches the entire genome for sequences containing the motifs of the four key sites in the U1 snRNA (Hamm et al., 1990; Kyriakopoulou et al., 2006): the 5' splice site complementary sequence (ACTTACC), the sequence forming loop I of the secondary structure that serves as the binding site of the protein U1-70K (GATCANGAAG), part of the sequence of the loop II that serves as the binding site of the protein U1-A (CATTGCAC) and the sequence of the Sm − site (ATTTNTG) . These positions have a high degree of conservation in the multiple alignment of the U1 sequences from Rfam. We used empirically derived minimum and maximum distances between every pair of adjacent sites, based on the sequences in the Rfam database, as constraints in the search. We obtained a large number of candidates, which were then analyzed with the Infernal package. With this procedure we were able to identify only one further U1 snRNA in *M. grisea*.

Subsequently, we tried to identify U1 snRNAs in the rest of species using the same motif search approach but relaxing some of the sequence constraints in the key motifs. Indeed, there are known cases, like *S. pombe* (Porter et al., 1990), where some of these positions diverge considerably compared to the majority of the species. We used several combinations of changes in the key motifs, and with these changes we were able to find the U1 snRNAs for *N. crassa, D. discoideum, C. parvum, D. hansenii, Y. lipolytica* and *A. fumigatus,* but not for the *U. maydis* and *C.neoformans*. The U1 snRNAs found in these species, as well as in *M. grisea*, were all similar to the metazoan U1 snRNAs.

Finally, we undertook a third approach in an attempt to locate the U1 snRNAs in the two remaining species, *U. maydis* and *C. neoformans*. We used the tool cmsearch from the Infernal package, which searches the entire genome for sequences that fit a secondary structure model built from the alignment of all known U1snRNAs in Rfam. To validate this approach, we applied this search to all the genomes above and verified that the U1 found was the same one as the one obtained previously. Using this tool we analyzed the genomes of *U. maydis* and *C. neoformans*. For *C. neoformans* we identified a possible candidate for U1 snRNA with low score, possibly due to the difference in length with the

other U1 snRNAs. For *U. maydis* we were unable to identify a good candidate. To eliminate the possibility that the U1 snRNAs in these two species are in fact more similar to the U1 in *S. cerevisiae*, we applied the script approach followed by the Infernal package analysis, as well as the cmsearch, both using the four yeast species *S. cerevisiae*, *C. glabrata*, *K. lactis*, and *E. gossypii*. However, this analysis did not yield any U1 snRNA candidates in these two organisms.

References

Abovich, N., Liao, X.C., and Rosbash, M. (1994). The yeast MUD2 protein: an interaction with PRP11 defines a bridge between commitment complexes and U2 snRNP addition. Genes & development *8*, 843-854.

Abril, J.F., Castelo, R., and Guigo, R. (2005). Comparison of splice sites in mammals and chicken. Genome Res *15*, 111-119.

Altschul, S.F., Gish, W., Miller, W., Myers, E.W., and Lipman, D.J. (1990). Basic local alignment search tool. J Mol Biol *215*, 403-410.

Ast, G. (2004). How did alternative splicing evolve? Nat Rev Genet *5*, 773-782.

Birney, E., Clamp, M., and Durbin, R. (2004). GeneWise and Genomewise. Genome Res *14*, 988-995.

Black, D.L. (2003). Mechanisms of alternative pre-messenger RNA splicing. Annual review of biochemistry *72*, 291-336.

Bouck, J., Fu, X.D., Skalka, A.M., and Katz, R.A. (1995). Genetic selection for balanced retroviral splicing: novel regulation involving the second step can be mediated by transitions in the polypyrimidine tract. Molecular and cellular biology *15*, 2663-2671.

Burge, C., and Karlin, S. (1997). Prediction of complete gene structures in human genomic DNA. J Mol Biol *268*, 78-94.

Carmel, I., Tal, S., Vig, I., and Ast, G. (2004). Comparative analysis detects dependencies among the 5' splice-site positions. Rna *10*, 828-840.

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.

Dou, Y., Fox-Walsh, K.L., Baldi, P.F., and Hertel, K.J. (2006). Genomic splice-site analysis reveals frequent alternative splicing close to the dominant splice site. Rna *12*, 2047-2056.

Du, H., and Rosbash, M. (2002). The U1 snRNP protein U1C recognizes the 5' splice site in the absence of base pairing. Nature *419*, 86-90.

Griffiths-Jones, S., Moxon, S., Marshall, M., Khanna, A., Eddy, S.R., and Bateman, A. (2005). Rfam: annotating non-coding RNAs in complete genomes. Nucleic Acids Res *33*, D121-124.

Hamm, J., Dathan, N.A., Scherly, D., and Mattaj, I.W. (1990). Multiple domains of U1 snRNA, including U1 specific protein binding sites, are required for splicing. The EMBO journal *9*, 1237-1244.

Kandels-Lewis, S., and Seraphin, B. (1993). Involvement of U6 snRNA in 5' splice site selection. Science (New York, N.Y *262*, 2035-2039.

Kent, O.A., Ritchie, D.B., and Macmillan, A.M. (2005). Characterization of a U2AF-independent commitment complex (E') in the mammalian spliceosome assembly pathway. Molecular and cellular biology *25*, 233-240.

Kielkopf, C.L., Lucke, S., and Green, M.R. (2004). U2AF homology motifs: protein recognition in the RRM world. Genes & development *18*, 1513-1526.

Kol, G., Lev-Maor, G., and Ast, G. (2005). Human-mouse comparative analysis reveals that branch-site plasticity contributes to splicing regulation. Hum Mol Genet *14*, 1559-1568.

Konarska, M.M., Vilardell, J., and Query, C.C. (2006). Repositioning of the reaction intermediate within the catalytic center of the spliceosome. Molecular cell *21*, 543-553.

Kretzner, L., Krol, A., and Rosbash, M. (1990). Saccharomyces cerevisiae U1 small nuclear RNA secondary structure contains both universal and yeast-specific domains. Proc Natl Acad Sci U S A *87*, 851-855.

Kupfer, D.M., Drabenstot, S.D., Buchanan, K.L., Lai, H., Zhu, H., Dyer, D.W., Roe, B.A., and Murphy, J.W. (2004). Introns and splicing elements of five diverse fungi. Eukaryot Cell *3*, 1088-1100.

Kyriakopoulou, C., Larsson, P., Liu, L., Schuster, J., Soderbom, F., Kirsebom, L.A., and Virtanen, A. (2006). U1-like snRNAs lacking complementarity to canonical 5' splice sites. Rna *12*, 1603-1611.

Lev-Maor, G., Sorek, R., Shomron, N., and Ast, G. (2003). The birth of an alternatively spliced exon: 3' splice-site selection in Alu exons. Science (New York, N.Y *300*, 1288-1291.

Lim, L.P., and Burge, C.B. (2001). A computational analysis of sequence features involved in recognition of short introns. Proc Natl Acad Sci U S A *98*, 11193-11198.

Long, M., de Souza, S.J., and Gilbert, W. (1997). The yeast splice site revisited: new exon consensus from genomic analysis. Cell *91*, 739-740.

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.

Manceau, V., Swenson, M., Le Caer, J.P., Sobel, A., Kielkopf, C.L., and Maucuer, A. (2006). Major phosphorylation of SF1 on adjacent Ser-Pro motifs enhances interaction with U2AF65. The FEBS journal *273*, 577-587.

Maris, C., Dominguez, C., and Allain, F.H. (2005). The RNA recognition motif, a plastic RNA-binding platform to regulate post-transcriptional gene expression. The FEBS journal *272*, 2118-2131.

Newman, A.J., and Norman, C. (1992). U5 snRNA interacts with exon sequences at 5' and 3' splice sites. Cell *68*, 743-754.

Notredame, C., Higgins, D.G., and Heringa, J. (2000). T-Coffee: A novel method for fast and accurate multiple sequence alignment. J Mol Biol *302*, 205-217.

Porter, G., Brennwald, P., and Wise, J.A. (1990). U1 small nuclear RNA from Schizosaccharomyces pombe has unique and conserved features and is encoded by an essential single-copy gene. Molecular and cellular biology *10*, 2874-2881.

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.

Rutz, B., and Seraphin, B. (1999). Transient interaction of BBP/ScSF1 and Mud2 with the splicing machinery affects the kinetics of spliceosome assembly. Rna *5*, 819-831.

Selenko, P., Gregorovic, G., Sprangers, R., Stier, G., Rhani, Z., Kramer, A., and Sattler, M. (2003). Structural basis for the molecular recognition between human splicing factors U2AF65 and SF1/mBBP. Molecular cell *11*, 965-976.

Sickmier, E.A., Frato, K.E., Shen, H., Paranawithana, S.R., Green, M.R., and Kielkopf, C.L. (2006). Structural basis for polypyrimidine tract recognition by the essential pre-mRNA splicing factor U2AF65. Molecular cell *23*, 49-59.

Slater, G.S., and Birney, E. (2005). Automated generation of heuristics for biological sequence comparison. BMC bioinformatics *6*, 31.

Smith, C.W., Chu, T.T., and Nadal-Ginard, B. (1993). Scanning and competition between AGs are involved in 3' splice site selection in mammalian introns. Molecular and cellular biology *13*, 4939-4952.

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.

Staley, J.P., and Guthrie, C. (1999). An RNA switch at the 5' splice site requires ATP and the DEAD box protein Prp28p. Molecular cell *3*, 55-64.

Thanaraj, T.A., and Robinson, A.J. (2000). Prediction of exact boundaries of exons. Briefings in bioinformatics *1*, 343-356.

Webb, C.J., and Wise, J.A. (2004). The splicing factor U2AF small subunit is functionally conserved between fission yeast and humans. Molecular and cellular biology *24*, 4229-4240.

Zamore, P.D., and Green, M.R. (1989). Identification, purification, and biochemical characterization of U2 small nuclear ribonucleoprotein auxiliary factor. Proc Natl Acad Sci U S A *86*, 9243-9247.

Supplementary figures

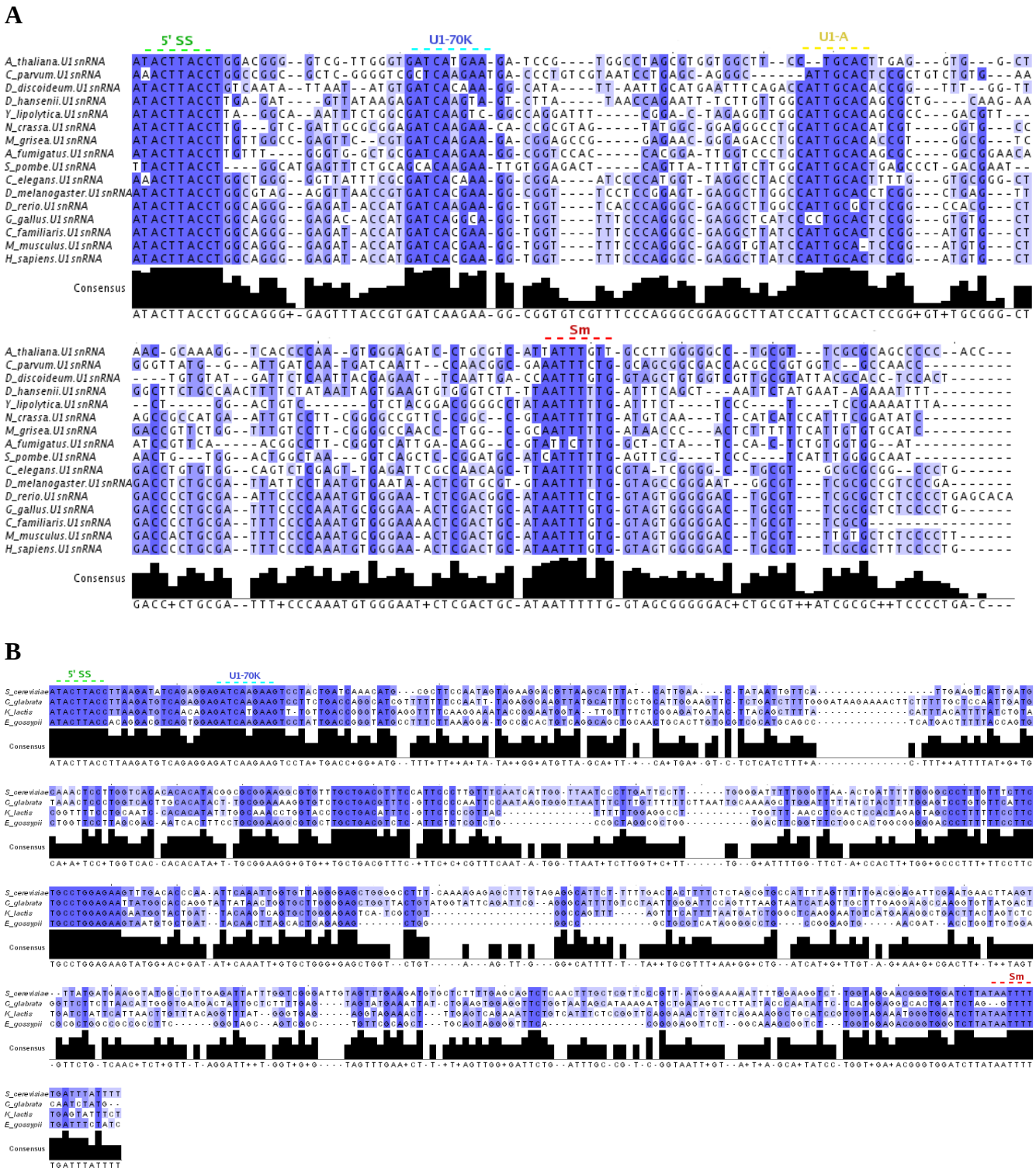

Figure S1: Full multiple sequence alignments of U1 snRNA molecules in metazoan-like **(A)** and *S. cerevisiae*-like **(B)** organisms. The binding sites to the 5'ss, U1-70K, to U1-A and to Sm proteins are marked.

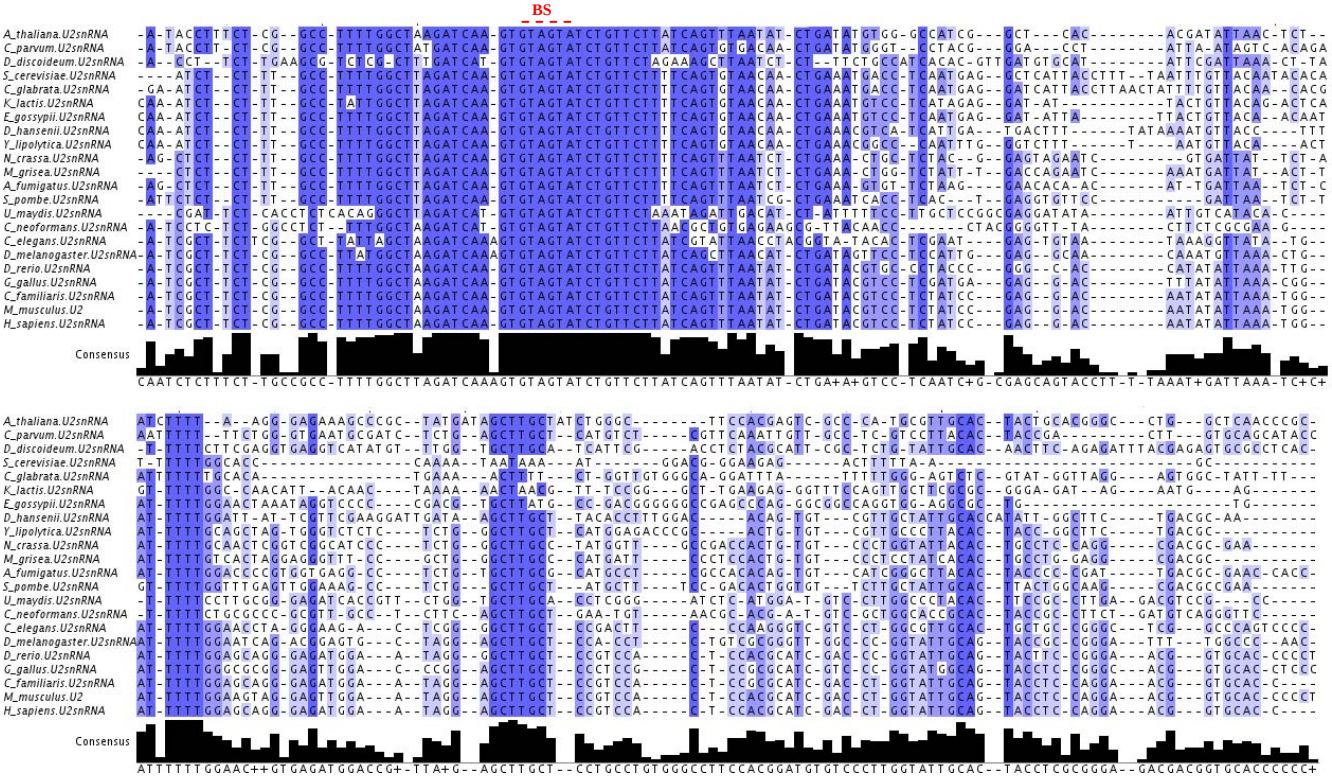

Figure S2: Full multiple sequence alignments of U2 snRNA molecules in the 22 organisms. The hexamer binding the BS is marked.

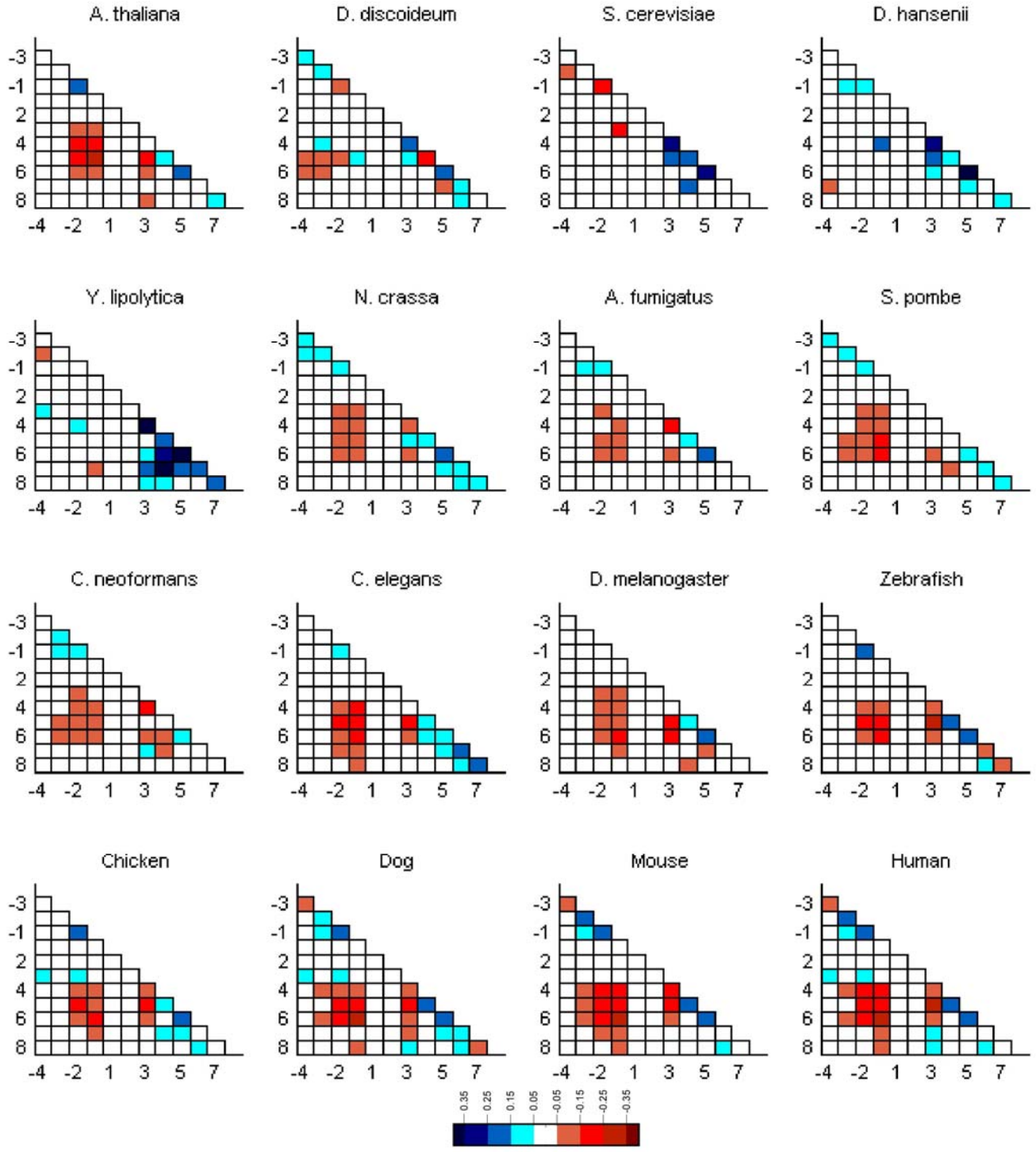

Figure S2

Figure S3: Dependency between pairs of positions of the 5'ss, shown for selected organisms. Positions -4 to -1 refer to positions within the exon, while positions 1 to 8 refer to positions within the intron. The Spearman rank correlation coefficient was calculated between each pair of positions and assigned a color-code according to its nature and strength, where different shades of blue and red represent different levels of positive and negative correlations, respectively. The exact range of the correlation coefficient represented by each shade can be viewed on the legend, at the bottom. Weak or statistically non-significant correlations ($p>0.01$) are presented in white.

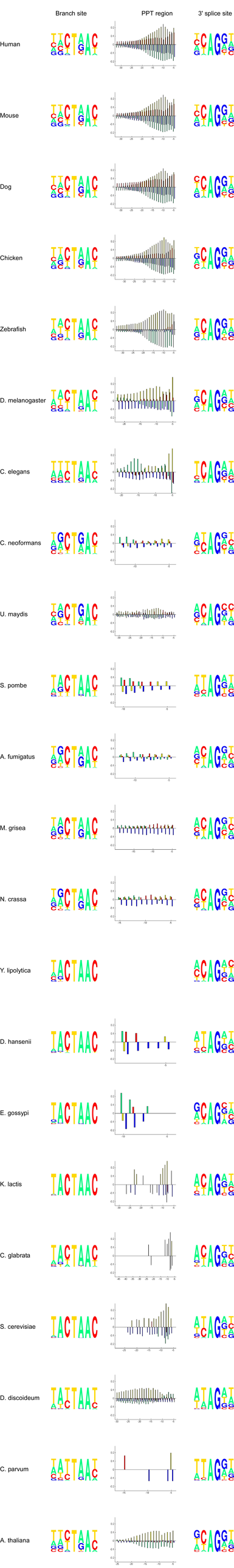
Branch site
PPT region
3' splice site
Human
Mouse
Dog
Chicken
Zebrafish
D. melanogaster
C. elegans
C. neoformans
U. maydis
S. pombe
A. fumigatus
M. grisea
N. crassa
Y. lipolytica
D. hansenii
E. gossypi
K. lactis
C. glabrata
S. cerevisiae
D. discoideum
C. parvum
A. thaliana

Figure S4: Downstream splicing signals among the 22 organisms. The left panel presents the BS motifs as in Figure 6A; the middle panel presents an overview of the PPT region, using bias plots as in Figure 3; the right panel presents the 3'ss in the form of pictograms. Note: In *Y. lipolytica*, the BS and the 3'ss form one consecutive stretch; therefore, this organism lacks a plot for the PPT region.

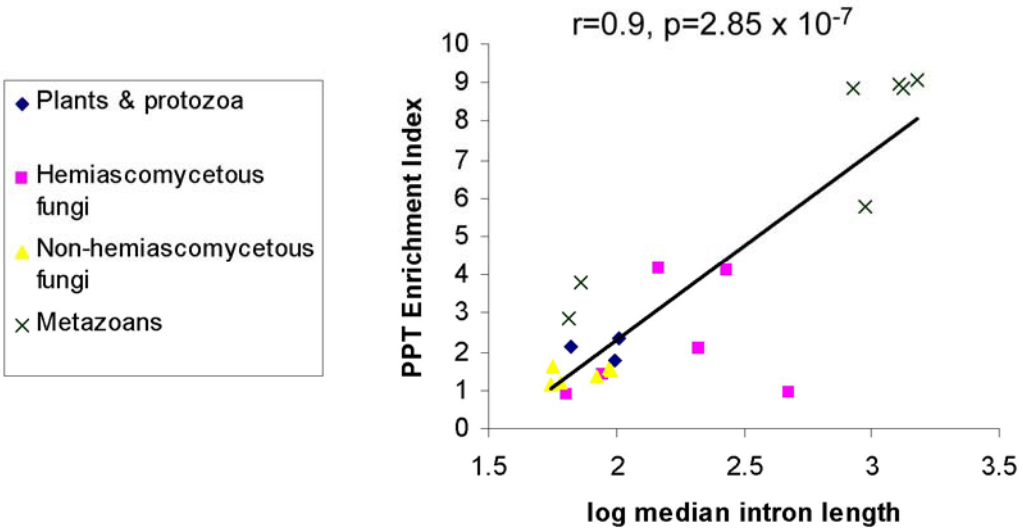

Figure S5: Correlation between median intron length and PPT enrichment index. The 22 organisms are divided into four groups, as shown by the legend to the left. A logarithmic scale to the base of 10 was used for the median intron length axis. The Pearson correlation and significance thereof is plotted at the top of the plot.

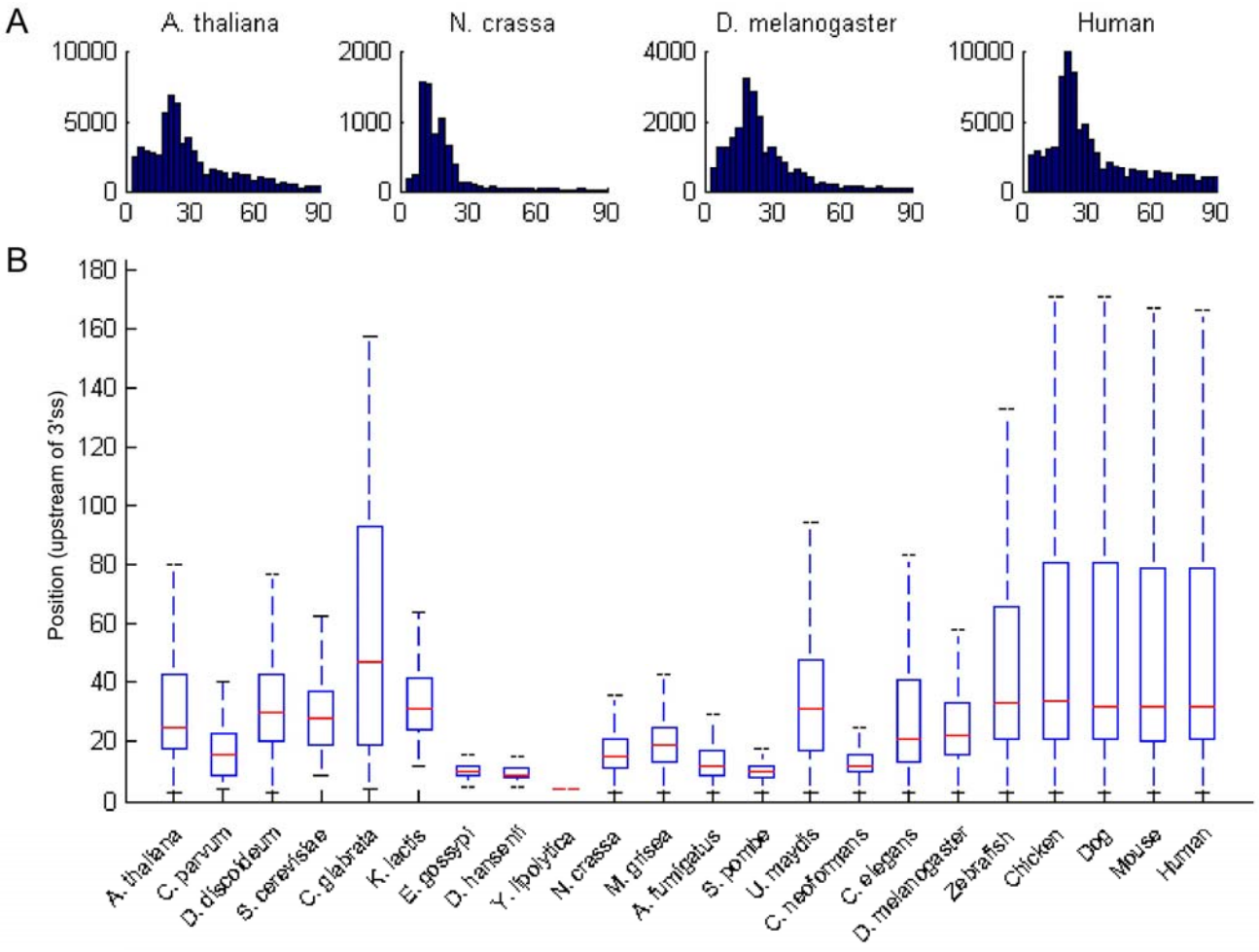

Figure S6: Distribution of the BS relative to the 3'ss. **(A)** Histograms of the distances between the first position of the BS and the 3'ss for sample organisms. The X axis represents the distance from the 3'ss, and the Y axis presents the number of occurrences **(B)** Boxplots of the distributions of the distances between the first position of the BS sequence and the 3'ss, for all organisms. The central (red) line marks the median; the rectangle marks the interquartile range, i.e. the range between the 25th and 75th percentile; the "whiskers" extend until 1.5 times the interquartile range or until the most extreme value of the distribution. Outlying values were not plotted.

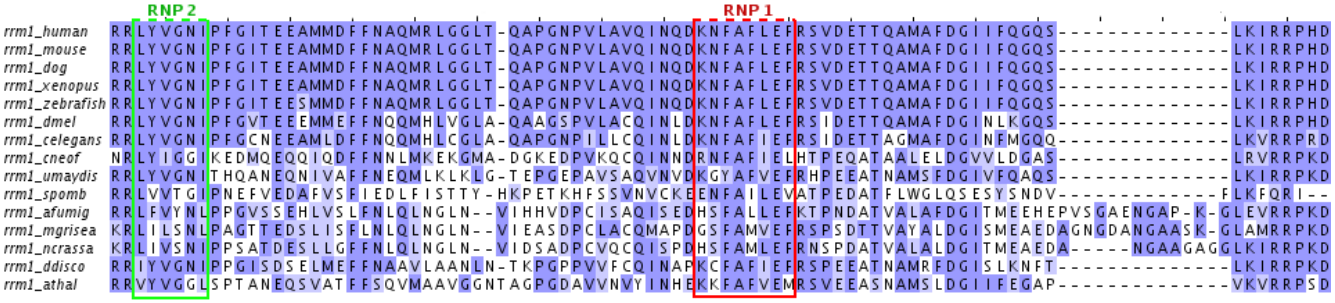

Figure S7. Multiple alignment of the RRM1 domain of the found U2AF65 homologs. The RNP1 and RNP2 motifs, relevant for RNA binding, are indicated. The darkness of the blue colour indicates the BLOSUM62 score.

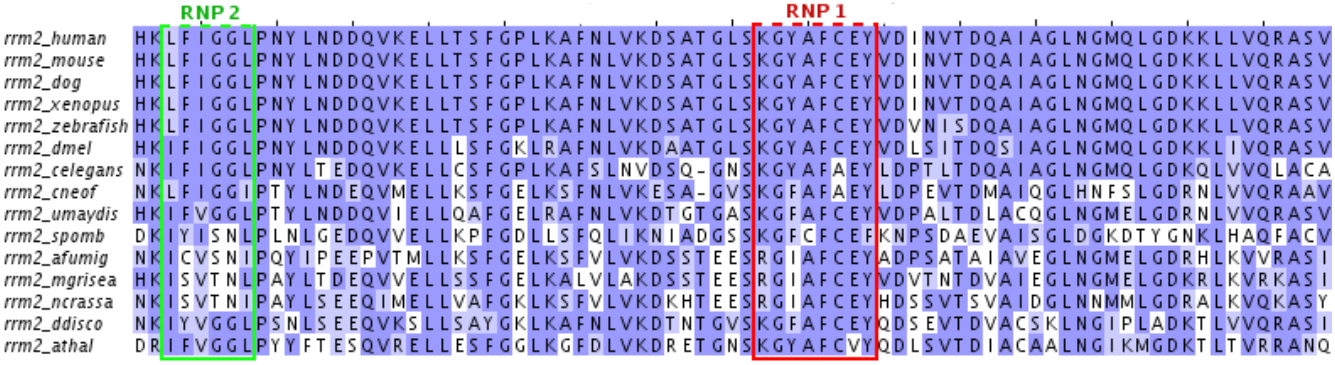

Figure S8. Multiple alignment of the RRM2 domain of the found U2AF65 homologs. The RNP1 and RNP2 motifs, relevant for RNA binding are indicated.

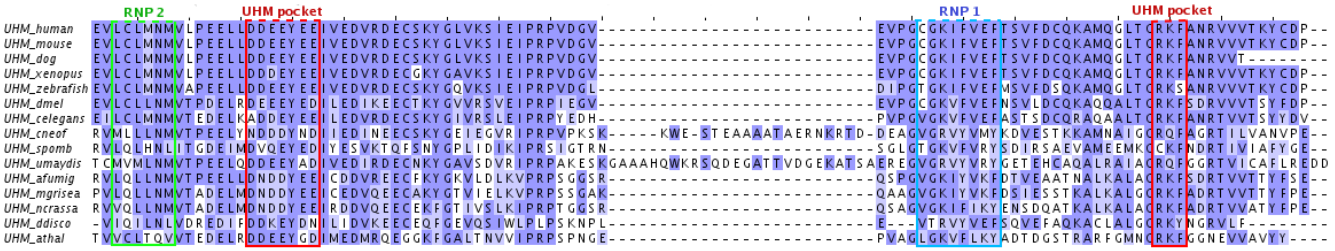

Figure S9. Multiple alignment of the UHM domain of the found U2AF65 homologs. The RNP1 and RNP2 motifs, relevant for RNA binding are indicated. The hydrophobic pocket that participates in the interaction with SF1, formed by two motifs, D-X-X-X-E and R-X-F, is also indicated.

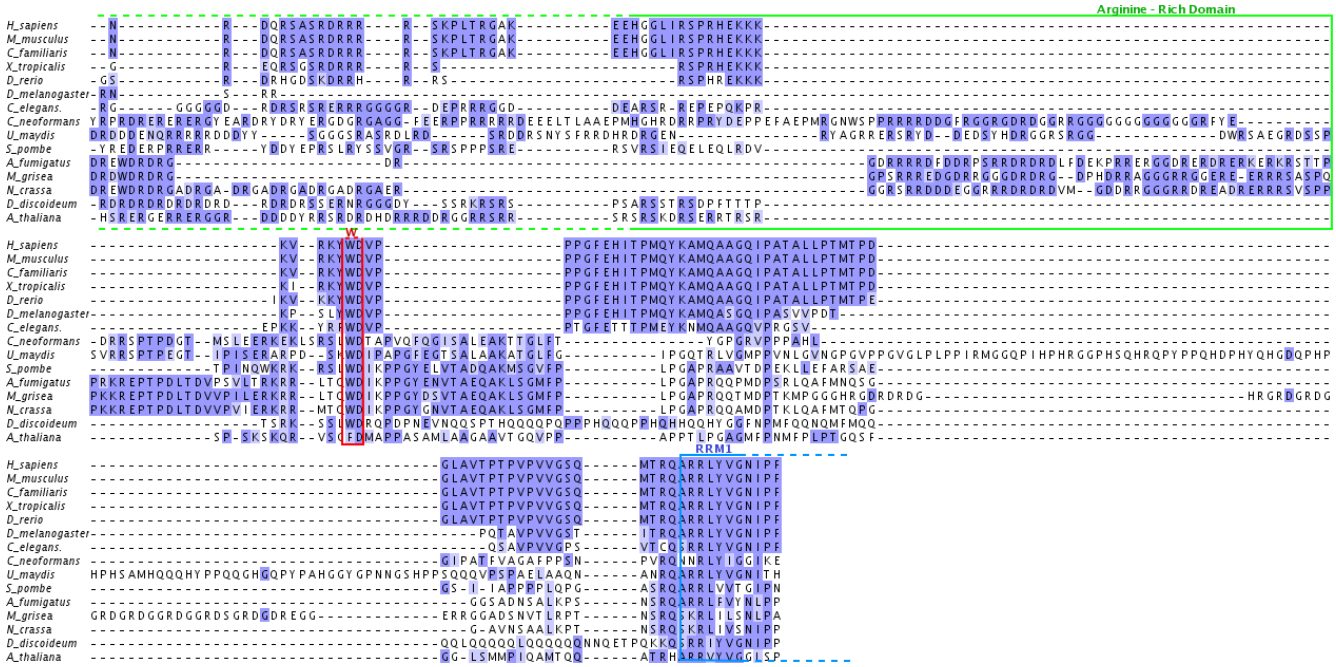

Figure S10. Multiple alignment of part of the N-terminal region of the U2AF65 homologs. The conserved tryptophan residue, or phenylalanine in the case of *A. thaliana*, is indicated. This residue is relevant for the interaction with U2AF35 and is located between the variable arginine-rich region and the RRM1 domain.

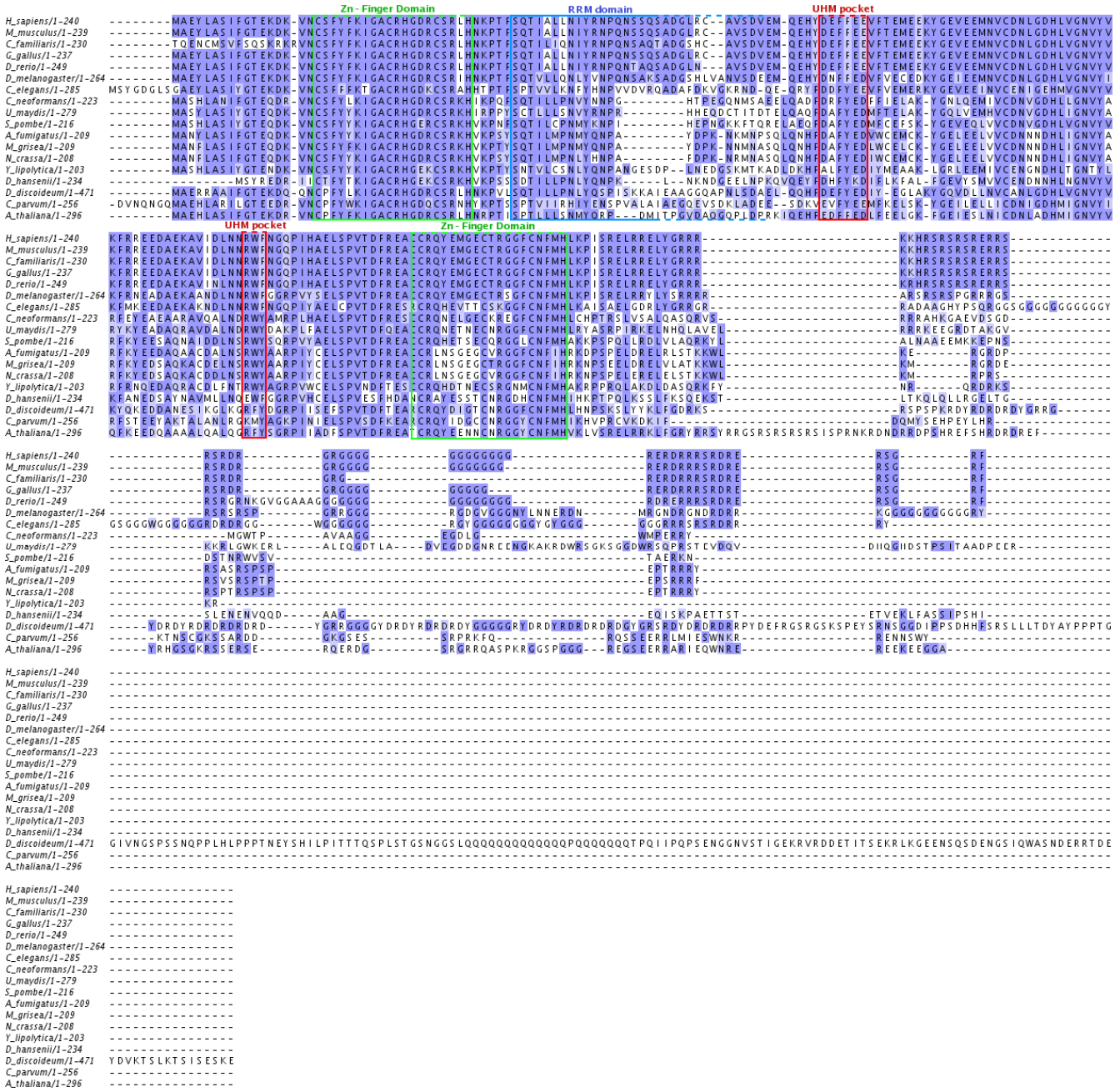

Figure S11. Multiple alignment of the U2AF35 homologs. The hydrophobic pocket in the UHM domain and the flanking zinc-finger domains are indicated.

Figure S12. Multiple alignment of SF1 homologs. The KH domain, the SPSP motif relevant for branch-site recognition, and the tryptophan relevant for the interaction with U2AF65 are indicated.

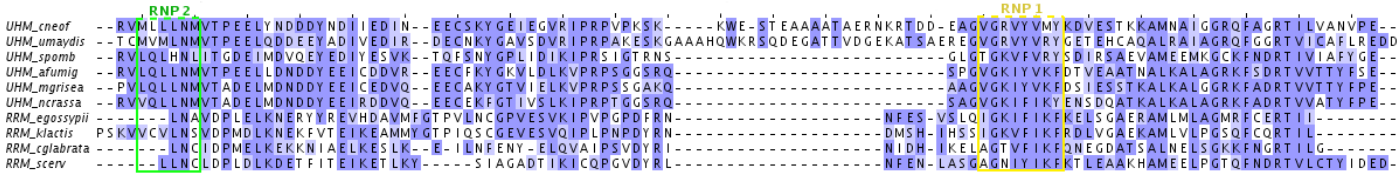

Figure S13. Multiple alignment of the UHM from U2AF65 homologs in fungal species and the MUD2 RRM in hemiascomycetous fungi. The conserved RNP1 and RNP2 motifs are highlighted.

Figure S14. Multiple alignment of the MSL5 sequences for *S. cerevisiae*, *C. glabrata*, *E. gossypii* and *K. lactis*. The KH domain and the SPSP motif are highlighted.

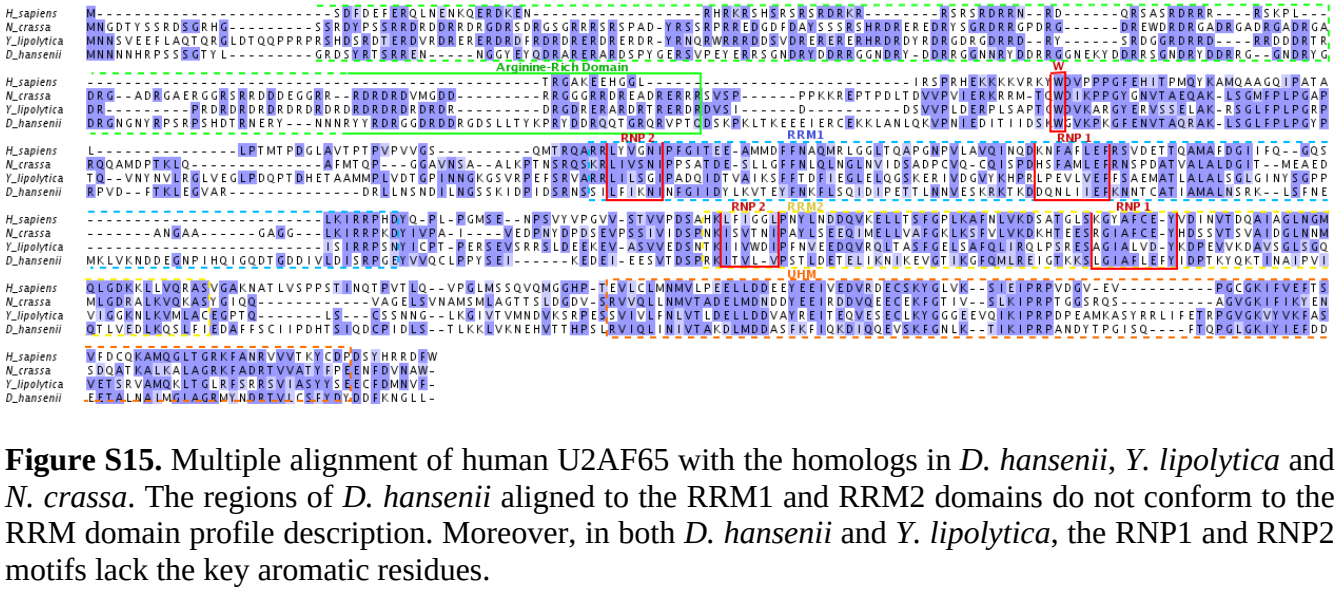

Figure S15. Multiple alignment of human U2AF65 with the homologs in *D. hansenii*, *Y. lipolytica* and *N. crassa*. The regions of *D. hansenii* aligned to the RRM1 and RRM2 domains do not conform to the RRM domain profile description. Moreover, in both *D. hansenii* and *Y. lipolytica*, the RNP1 and RNP2 motifs lack the key aromatic residues.

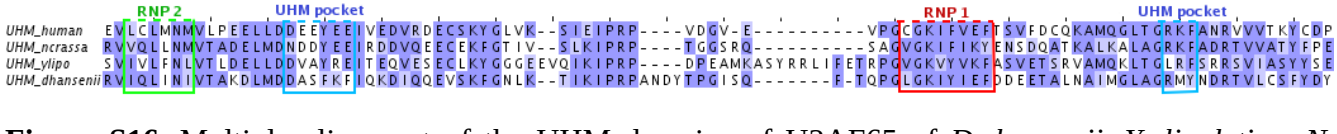

Figure S16. Multiple alignment of the UHM domains of U2AF65 of *D. hansenii*, *Y. lipolytica*, *N. crassa*, and human. The position of the hydrophobic pocket is indicated.

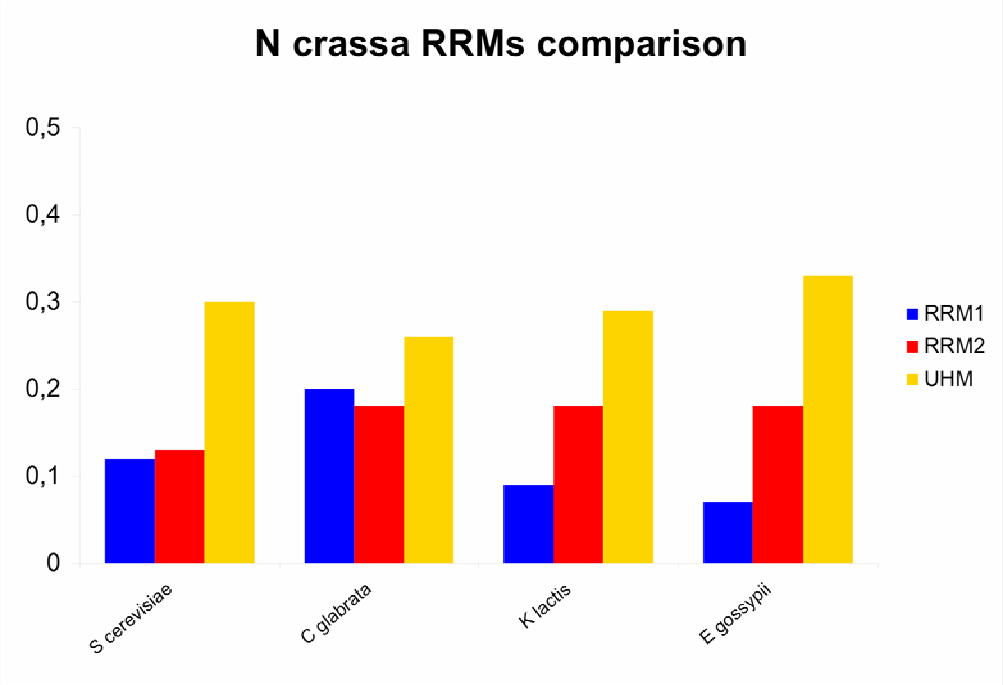

Figure S17. Comparison of the three RRM domains of U2AF65 with the RRM of MUD2. The plot shows the sequence similarity between each of the U2AF65 RRM domains (RRM1, RRM2 and UHM) of *N. crassa* with the RRM of MUD2 in *S. cerevisiae*, *K. lactis*, *C. glabrata* and *E. gossypii*. The same trends are found for the U2AF65 RRM domains in other species (data not shown).

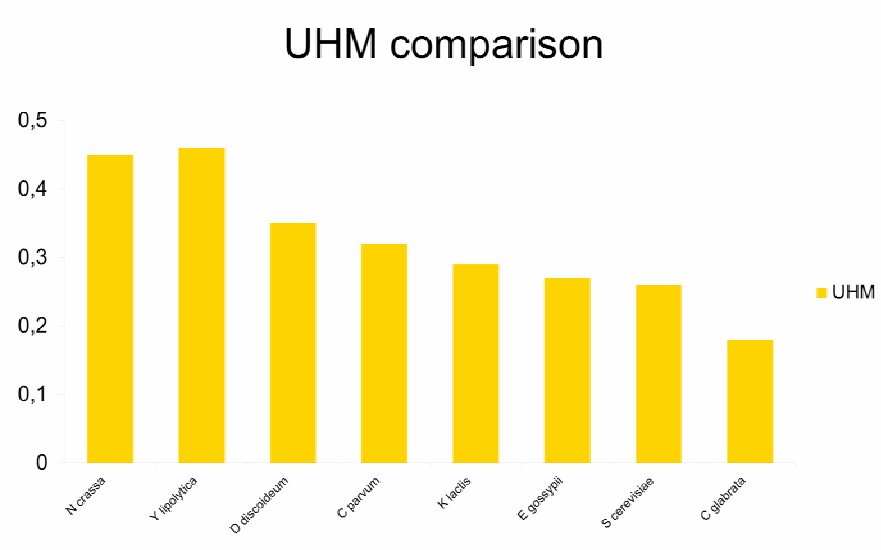

Figure S18. Comparison of the UHM domain of *C. parvum* with other species. The plot shows the sequence similarity between the UHM domain in human U2AF65 and the UHM domains of *N. crassa*, *Y. lipolytica* and *C. parvum*, and the RRM domain of MUD2 in *E. gossypii*, *S. cerevisiae* and *C. glabrata*.

Supplementary Tables

	Database compilation						Intronic nucleotide composition				BS analysis	
	Extracted introns	length < 15	Non-canonical	GC introns	U12 introns	Introns in database	A	T	C	G	% introns with BS	PPT distance
A. thaliana	115616	32	92	932	274	114286	0.28	0.40	0.15	0.17	0.57	25
C. parvum	44	0	0	0	0	44	0.34	0.41	0.11	0.13	0.68	15.5
D. discoideum	17377	28	15	2	6	17326	0.42	0.46	0.06	0.06	0.62	30
S. cerevisiae	314	45	2	8	1	258	0.33	0.34	0.16	0.17	0.59	28
C. glabrata	84	2	2	0	0	80	0.33	0.34	0.15	0.18	0.34	47
K. lactis	130	1	2	0	0	127	0.33	0.33	0.16	0.18	0.54	31
E. gossypi	228	7	1	2	0	218	0.24	0.22	0.24	0.30	0.94	10
D. hansenii	357	0	10	1	0	346	0.35	0.31	0.15	0.20	0.87	9
Y. lipolytica	739	2	14	2	0	721	0.26	0.25	0.24	0.25	0.98	4
N. crassa	10025	0	0	0	2	10023	0.25	0.27	0.26	0.22	0.77	15
M. grisea	19466	0	0	0	3	19463	0.26	0.28	0.24	0.22	0.73	19
A. fumigatus	18322	47	25	33	5	18212	0.25	0.29	0.24	0.22	0.78	12
S. pombe	4765	10	8	5	8	4734	0.31	0.39	0.14	0.16	0.88	10
U. maydis	4882	4	0	0	0	4878	0.23	0.27	0.26	0.24	0.51	31
C. neoformans	34336	14	14	566	2	33740	0.25	0.31	0.21	0.22	0.78	12
C. elegans	99251	4	127	387	38	98695	0.33	0.34	0.16	0.16	0.51	21
D. melanogaster	41429	0	46	229	9	41145	0.29	0.31	0.20	0.20	0.58	22
Zebrafish	200924	0	2361	3912	430	194221	0.32	0.33	0.18	0.18	0.52	33
Chicken	171192	0	1274	1970	322	167626	0.28	0.31	0.19	0.21	0.55	34
Dog	176104	0	366	3243	436	172059	0.28	0.31	0.20	0.21	0.56	32
Mouse	181599	1	1685	1685	462	177766	0.27	0.30	0.21	0.22	0.58	32
Human	187740	0	1292	1812	491	184145	0.28	0.31	0.20	0.21	0.56	32

Table S1. General statistics on database compilation, intron characteristics, and branch site analysis. Database compilation data contains the full number of extracted introns based on GenBank annotations (after discarding alternatively spliced isoforms) and the number of introns filtrated due to considerations of length, non-canonical splicing borders, and following U12 intron filtration (see Methods). The background frequency of each of the four nucleotides is shown, based on which information content of the splicing signals was calculated and the positional bias plots were derived. Branch site analysis data includes the percentage of introns in which BSs were detected in each organism and the median distance between the termination of the BS and the 3'ss.

	PPT									
	Enrichment Index	p value (Mann-Whitney)	% Introns	Mean Lengths	T:C Ratio	T:C Adjusted Ratio	p value (χ2)	G:A ratio	G:A Adjusted Ratio	p value (χ2)
A. thaliana	1.79	0	32.1	13.01	2.88	1.11	0	1.07	1.74	0
C. parvum	2.15	0.045	30.0	13.33	5.38	1.39	0.11	0.15	0.39	0.20
D. discoideum	2.33	0	29.6	21.51	19.25	2.54	0	0.05	0.33	0
S. cerevisiae	4.19	0	40.1	14.20	3.95	1.90	0	0.34	0.64	0.05
C. glabrata	0.92	0.972	7.4	10.17	2.63	1.14	0.65	0.00	0.00	0.21
K. lactis	4.14	2.6E-11	35.3	13.28	2.43	1.22	0.06	0.48	0.87	0.66
E. gossypi	0.91	0.858	3.9	7.80	0.70	0.74	0.21	0.50	0.41	0.45
D. hansenii	1.42	0.129	5.6	6.94	6.31	2.96	2.4E-05	0.00	0.00	0.45
Y. lipolytica	2.08	0.002	4.5	11.91	1.18	1.12	0.28	0.80	0.85	0.68
N. crassa	1.35	1.7E-12	20.7	11.06	1.08	1.01	0.32	0.83	0.93	0.16
M. grisea	1.57	0	20.8	11.15	1.26	1.11	0	0.92	1.07	0.07
A. fumigatus	1.18	2.0E-06	15.2	9.82	1.29	1.06	1.3E-05	0.87	0.98	0.64
S. pombe	1.63	4.4E-16	16.1	9.22	3.15	1.10	0.001	0.37	0.72	0.009
U. maydis	1.49	1.8E-05	20.3	14.33	1.28	1.24	0	1.41	1.33	7.7E-07
C. neoformans	1.14	1.5E-05	15.1	10.21	1.47	1.02	0.02	0.66	0.76	1.9E-14
C. elegans	2.87	0	35.4	10.09	3.25	1.51	0	0.29	0.61	0
D. melanogaster	3.80	0	47.9	12.80	1.78	1.18	0	0.62	0.93	5.7E-07
Zebrafish	5.76	0	54.1	14.84	2.18	1.16	0	1.22	2.16	0
Chicken	8.84	0	68.0	15.74	1.71	1.07	0	1.23	1.59	0
Dog	8.97	0	71.4	16.08	1.33	0.88	0	1.18	1.55	0
Mouse	8.86	0	73.0	15.81	1.41	0.99	0	1.22	1.54	0
Human	9.08	0	72.8	16.02	1.41	0.94	0	1.09	1.44	0

Table S2. Statistics pertaining to the PPT analysis. The PPT enrichment index, its statistical significance, the percentage of introns in which PPTs were found and the mean lengths of the identified PPTs are plotted for each organism. The full results for nucleotide composition analyses of the PPT are shown as well. These include the absolute T:C ratio in the PPT, the T:C ratio after adjustment for the background T:C ratio and the statistical significance thereof, with identical analyses for the G:A ratio.

Organism	Annotation	Gene ID
A. thaliana	Arabidopsis thaliana U1 snRNA gene, complete sequence	AY222070.1/1-157
C. parvum		NC_006987.1.57300-57460
D. discoideum		DDB0232429.6472256-6472422
S. cerevisiae	Yeast (S.cerevisiae) U1 (snR19) RNA gene	M17205.1/245-812
C. glabrata		CR380957.1/492012-492606
K. lactis	Kluyveromyces lactis U1 small nuclear RNA	U03475.1/1-528
E. gossypi		AE016815.2/55081-54598
D. hansenii		Deha0A.648183-648344
Y. lipolytica		Yali0B.1936777-1936917
N. crassa		contig_7.73.54427-54581
M. grisea	M_grisea supercontig_5.187 s 1311800 e 1312017 o -	M_grisea supercontig_5.187
A. fumigatus	Supercontig 98 chr_1 AAHF01000007	Supercontig 98 chr_1 AAHF01000007
S. pombe	S.pombe U1 small nuclear RNA gene (snu1), complete cds	M29062.1/238-387
C. elegans	Caenorhabditis elegans DNA encoding U1-1 snRNA	X51371.1/181-345
D. melanogaster	D. melanogaster U1 small nuclear RNA	K00787.1/2-165
zebrafish		AL929029.7/75931-75768
chicken	chicken u1 small nuclear rna (snrna).	J00914.1/146-309
dog	Canis familiaris U1 snRNA gene	L33345
mouse	mouse u1 small nuclear rna (snrna) gene.	J00645.1/51-213
human	Homo sapiens U1 snRNA gene	V00591.1/394-557

Table S3. Gene IDs and annotations (when available) of the extracted U1 snRNA genes are presented.

Organism	Annotation	Gene ID
A. thaliana		AC004138.3/33098-33293
1C. parvum		Contig: NC_006984.1 Coordinates: 933830-934289
D. discoideum	Dictyostelium discoideum u2 snRNA, clone ddR-19	AJ699380.1/1-206
S. cerevisiae	Yeast (S.cerevisiae) LSR1 gene encoding the yeast homolog of vertebrate U2 small nuclear RNA	M14625.1/328-521
C. glabrata		CR380957.1/704696-704866
K. lactis		CR382126.1/991021-991198
E. gossypi		AE016819.2/408896-409052
D. hansenii		CR382136.1/895858-896054
Y. lipolytica		CR382129.1/446178-446370
N. crassa		BX294012.1/95728-95534
M. grisea		Supercontig 5.178 Coordinates: 755070-755524
A. fumigatus		AL683874.1/16263-16072
S. pombe	S. pombe snu2 gene for U2 snRNA	X55772.1/223-413
U. maydis		Contig: 1.91 Coordinates: 4658-5085
C. neoformans		AE017345.1/926592-926777
C. elegans	Caenorhabditis elegans DNA encoding U2-9 snRNA	X51381.1/239-429
D. melanogaster	Drosophila melanogaster gene for small nuclear U2 RNA (clone 131B)	X04243.1/69-264
Zebrafish		BX005336.10/96115-95925
Chicken	Chicken U2 small nuclear RNA gene	M12856.1/361-551
Dog		Chromosome: 14 Sequence: 47884380-47884966
Mouse	Mouse U2 snRNA gene	X07913.1/1061-1251
Human	Human gene for small nuclear RNA U2	X01408.1/259-449

Table S4. Gene IDs and annotations (when available) of the extracted U2 snRNA genes are presented.

Organism	**Annotation**	**Gene ID**
Homo sapiens	U2AF2 - Chromosome 19 at location .	ENSG00000063244
Mus musculus	U2AF2 - Chromosome 7 at location .	ENSMUSG00000030435
Canis familiaris	U2AF2 - Chromosome 1 at location .	ENSCAFG00000002551
Xenopus tropicalis	U2AF2 - Scaffold_356 at location	ENSXETG00000019128
Danio rerio	Hypothetical protein LOC557103 - Chromosome 16 at location .	ENSDARG00000012505
Drosophila melanogaster	U2AF 50 Kda - Chromosome X at location	CG9998
Caenorhabditis elegans	UAF-1 - Chromosome III at location .	Y92C3B.2
Arabidopsis thaliana	ATU2AF65A – LOCUS AT4G36690 – Gene model AT4g36690.1	Gene:2115279
Schizosaccharomyces pombe	U2AF 59 Kda – Chromosome 2 at location 993555 – 1035108	Prp2
Cryptococcus neoformans JEC21	cn-jec21_chr6	cneo_JEC21_TIGR:CNF01250
Neurospora crassa	Conserved hypothetical protein – Chromosome I - Contig 7: 59978-62026 +	NCU03039.3
Aspergillus fumigatus	U2AF large subunit - AFU293 – Chromosome 7	AFUA_7G05310
Ustilago maydis	Hypothetical protein – umay_1 Contig 192: 99068-101251 -	UM05363.1
Magnaporthe grisea	Hypothetical protein – mgri_2.78	BRD:MG00348.4
Dictyostelium discoideum	U2AF2 on chromosome: 4 position 4405482 to 4407497	DDB0186947
Yarrowia lipolytica	Possible U2AF65 - Deha0G:352941..354956	DEHA0G04609g
Debaryomyces hansenii	Possible U2AF65 - Yali0E:1215043..1216848	YALI0E09889g
Cryptosporidium parvum	Splicing factor U2AF 3 RRMs - Chromosome6, positions 388,760 to 390,238	cgd6_1680
Cryptosporidium parvum	Splicing factor – like protein, putative RRM -Chromosome2, positions 319,204 to 320,520	cgd2_1480
Trichomonas vaginalis	Coordinates (5'-3') 123584 - 122382 on assembly	TVAG_453940

Table S5. List of the identified U2AF65 homologs. The gene ids, genome position and annotations from the genome project (when available) are provided. The two hemiascomycetous fungi *Y. lipolytica* and *D. hansenii*, and the protozoan *C. parvum*, are included.

Organism	**Annotation**	**Gene ID**
Homo sapiens	U2AF1 - Chromosome 21 at location .	ENSG00000160201
Mus musculus	U2AF1 - Chromosome 17 at location	ENSMUSG00000061613
Canis familiaris	U2AF35 - Chromosome 31 at location	ENSCAFG00000010572
Gallus gallus	U2AF35 - Chromosome 1 at location	ENSGALG00000016198
Xenopus tropicalis		ENSXETG00000004456
Danio rerio	U2AF1 - Chromosome 9 at location .	ENSDARG00000015325
Drosophila melanogaster	Splicing factor U2af 38 kDa subunit - Chromosome 2L at location .	CG3582
Caenorhabditis elegans	uaf-2 encodes an essential U2AF35 homolog clustered in an operon with cyp-13 (RRM/cyclophilin) - Chromosome IV at location .	Y116A8C.35
Arabidopsis thaliana	AtU2AF35a - The atU2AF35a protein and its homolog, atU2AF35b, contain most of the conserved domains of hsU2AF35	AT1G27650.1
Schizosaccharomyces pombe	U2AF-23 - Chromosome 1 Contig Location 5325611..5326261	SPAP8A3.06
Cryptococcus neoformans JEC21	U2AF35 – putative protein Chromosome: 9 Coordinates (5' - 3'):429055 - 428109 on assembly.	CNI01460
Neurospora crassa	Splicing factor U2AF 23 kDa subunit -Contig 9: 208083-209123 +	NCU03261
Aspergillus fumigatus	U2 auxiliary factor small subunit, putative – AFU293 chr_3 5285-5132	AFUA_3G02380
Ustilago maydis	U2Af35 homolog -hypothetical protein Contig 132: 72236-73075 +	
Magnaporthe grisea	U2Af35 hypothetical protein similar to (NCU03261.1) - Contig 2.1897 16727-17530.	MG09948.4
Yarrowia lipolytica	U2Af35 homolog - orf	YALI0F06292g
Debaryomyces hansenii	U2Af35 homolog – orf	DEHA0A03531g
Dictyostelium discoideum	U2Af35 homolog – Chr. 4 Contig Location 2804013..2806191 Length: 2179 bp	DDB0218665
Cryptosporidium parvum	U2AF35 putative protein - Chromosome8, positions 1,290,722 to 1,291,492	cgd8_5240
Trichomonas vaginalis	Coordinates (5'-3') 15901 - 16626 on assembly	TVAG_171620

Table S6. List of the identified U2AF35 homologs. The gene ids, genome position and annotations from the genome project (when available) are provided.

Organism	**Annotation**	**Gene ID**
Homo sapiens	SF1 - Chromosome 11 at location	ENSG00000168066
Mus musculus	SF1 - Chromosome 19 at location .	ENSMUSG00000024949
Canis familiaris	SF1 - Chromosome 18 at location .	ENSCAFG00000014239
Xenopus tropicalis	SF1 - Scaffold_146 at location .	ENSXETG00000008434
Danio rerio	SF1 - Chromosome 7 at location .	ENSDARG00000008188
Danio rerio	SF1- Chromosome 7 at location . Different gene coding for the same protein with the same sequence.	ENSDARG00000055095
Drosophila melanogaster	SF1 - Chromosome 3R at location .	CG5836
Caenorhabditis elegans	Ortholog of SF1 – Chromosome IV at location	Y116A8C.32
Arabidopsis thaliana	similar to Splicing factor 1/branch point binding protein – Chromosome 5, map – 20866123 - 20869560 bp	AT5G51300.1
Schizosaccharomyces pombe	Sf1 - Chromosome 3 Contig Location complement 556396..558205	SPCC962.06c (bpb1)
Cryptococcus neoformans JEC21	Splicing factor SF1, putative - Chromosome: 4 Coordinates (5' - 3'):778648 - 780590	163.m02728 (CND02880)
Neurospora crassa	NCU04110: hypothetical protein similar to zinc knuckle transcription factor Zfm1 – Chromosome V Contig 13: 789776-791985 -	NCU04110.3
Aspergillus fumigatus	SF1 - zinc knuckle transcription factor/splicing factor MSL5/ZFM1, putative - Chromosome: 3 Coordinates (5' - 3'):1257588 - 1255294 on assembly	AFUA_3G10840
Ustilago maydis	SF1 - UM06386: hypothetical protein - Contig 247: 93691-95568 +	
Magnaporthe grisea	SF1- hypothetical protein - contig 2.1252 – 32897 – 35268 +	
Yarrowia lipolytica	SF1/MSL5 - ORF from sense	YALI0F18370g
Debaryomyces hansenii	SF1/MSL5 - ORF from sense	DEHA0D07975g
Saccharomyces cerevisiae	MSL5 - chrXII, positions 370,823 to 392,253	YLR116W
Kluyveromyces lactis	MSL5 - ORF from antisense	KLLA0F18018g
Candida glabrata	MSL5 - ORF from sense	CAGL0D02354g
Ashbya gossypii	MSL5 - Chromosome VII at location 352,996-354,519.	AGL183C
Dictyostelium discoideum	SF1 - Chromosome 6 Contig Location 3023424..3025137 length: 1714 bp (sequence contains several N)	DDB0192004
Cryptosporidium parvum	SF1 putative - Chromosome4, positions 309,668 to 311,083	cgd4_1210
Trichomonas vaginalis	Coordinates (5'-3') 17731 - 16778 on assembly 1047229023758 Locus :92181.m00200	TVAG_343950

Table S7. List of the identified SF1 homologs. The gene ids, genome position and annotations from the genome project (when available) are provided. The MSL5 homologs in the hemiascomycetous species are included as well.

Organism	**Annotation**	**Gene ID**
Saccharomyces cerevisiae	YEAST Splicing factor MUD2	YKL074C
Candida glabrata	Cagl0L:569779..571320	CAGL0L05038g
Ashbya gossypii	Possible Mud2 homolog – Chromosome IV at location 938,152-939,153	ADR130W
Kluyveromyces lactis	Possible Mud2 homolog - Klla0F:2448748..2450517	KLLA0F26433g

Table S8. List of the identified MUD2 homologs. The gene ids, genome position and annotations from the genome project (when available) are provided.