Published online 12 August 2006 *Nucleic Acids Research, 2006, Vol. 34, No. 14* ***3955–3967***
doi:10.1093/nar/gkl556

Comprehensive splice-site analysis using comparative genomics

Nihar Sheth, Xavier Roca, Michelle L. Hastings, Ted Roeder, Adrian R. Krainer and Ravi Sachidanandam*

Cold Spring Harbor Laboratory, 1 Bungtown Road, Cold Spring Harbor, NY 11724, USA

Received June 7, 2006; Revised July 13, 2006; Accepted July 17, 2006

ABSTRACT

We have collected over half a million splice sites from five species—*Homo sapiens, Mus musculus, Drosophila melanogaster, Caenorhabditis elegans* and *Arabidopsis thaliana*—and classified them into four subtypes: U2-type GT–AG and GC–AG and U12-type GT–AG and AT–AC. We have also found new examples of rare splice-site categories, such as U12-type introns without canonical borders, and U2-dependent AT–AC introns. The splice-site sequences and several tools to explore them are available on a public website (SpliceRack). For the U12-type introns, we find several features conserved across species, as well as a clustering of these introns on genes. Using the information content of the splice-site motifs, and the phylogenetic distance between them, we identify: (i) a higher degree of conservation in the exonic portion of the U2-type splice sites in more complex organisms; (ii) conservation of exonic nucleotides for U12-type splice sites; (iii) divergent evolution of *C.elegans* 3′ splice sites (3′ss) and (iv) distinct evolutionary histories of 5′ and 3′ss. Our study proves that the identification of broad patterns in naturally-occurring splice sites, through the analysis of genomic datasets, provides mechanistic and evolutionary insights into pre-mRNA splicing.

INTRODUCTION

Pre-mRNA splicing is a nuclear process that is conserved across eukaryotes (1). This process involves the recognition of exon–intron junctions by the spliceosome, and intron excision through a two-step transesterification reaction (2,3). The spliceosome is a large and dynamic complex assembled from RNA and protein components, including four small nuclear RNAs (snRNAs) and associated proteins that make up the small nuclear ribonucleoprotein particles (snRNPs) (4). The spliceosome recognizes three conserved sequences at or near the exon–intron boundaries, namely the 5′ splice site (5′ss), the branch point sequence (BPS) and the 3′ss (Figure 1). In the first catalytic step, the 5′ end of the intron is cleaved and covalently joined to an Adenosine (A) on the BPS. In the second catalytic step, the neighboring exons are joined and the intron is excised as a lariat. There are at least two classes of pre-mRNA introns, based on the splicing machineries that catalyze the reaction:

- U2 snRNP-dependent introns make up the majority of all introns and are excised by spliceosomes containing the U1, U2, U4, U5 and U6 snRNPs. These introns consist of three subtypes, according to their terminal dinucleotides: GT–AG, GC–AG and AT–AC introns.
- U12 snRNP-dependent introns are the minor class of introns and are excised by spliceosomes containing U11, U12, U4atac, U6atac and U5 snRNPs. These introns mainly consist of two subtypes, as defined by their terminal dinucleotides: AT–AC and GT–AG introns. In addition, a small fraction of the U12-type introns exhibit other terminal dinucleotides (5–7).

Whereas U2-type introns have been found in virtually all eukaryotes (1) and comprise the vast majority of the splice sites found in any organism, U12-type introns have only been identified in vertebrates, insects, jellyfish and plants (8).

U2-type intron splicing initially involves base pairing of U1 snRNA to the 5′ss and U2 snRNA to the BPS (2) (Figure 1). The base pairing of U2 snRNA to the BPS is facilitated by the binding of the large subunit of the U2 Auxiliary Factor (U2AF65) to the poly-pyrimidine tract (PPT) located immediately upstream of the intron 3′ terminus, and binding of the small subunit (U2AF35) to the 3′ terminal AG dinucleotide of the intron (9,10). Following the initial recognition of the splice sites by the U1 and U2 snRNPs, the U4/U6/U5 tri-snRNP is recruited to the splice site leading to the two catalytic steps of splicing (11).

*To whom correspondence should be addressed. Tel: +1 516 367 8864; Fax: +1 516 367 8389; Email: sachidan@cshl.edu
Present address:
Nihar Sheth, Center for the Study of Biological Complexity, Virginia Commonwealth University, Richmond, VA 23284-2030, USA

The authors wish it to be known that, in their opinion, the first two authors should be regarded as joint First Authors

Figure 1. Initial steps in splice-site selection. The dark bars stand for exons, while the lines represent introns. ss stands for splice sites. AG is the 3′ terminus of the intron, where Y is a pyrimidine. PPT is the poly-pyrimidine tract. BPS is the branch point sequence, U1, U2, U11 and U12 are snRNPs. U2AF65 and U2AF35 are protein splicing factors. See text for details.

In U12-type introns, the roles of U1, U2, U4 and U6 snRNPs in U2-type introns are replaced by the U11, U12, U4atac and U6atac snRNPs, respectively (12–14). The overall similarity in the predicted secondary structure between analogous U2- and U12-type snRNAs suggests that the spliceosome rearrangements during catalysis are conserved between the two spliceosomes (15,16).

The 5′ss, 3′ss and BPS elements conform to specific consensus sequences as determined by the alignment of splice-site compilations (8,17–24). The U2-type splicing signals have highly degenerate sequence motifs; many different sequences can function as U2-type splice sites. In contrast, U12-type 5′ss and the BPS, which lies close to the 3′ end of the intron, are highly conserved (22,25). In most U2-type cases, the PPT is located immediately upstream of the AG but there are examples in alternatively spliced exons in *Drosophila melanogaster* with a longer PPT–AG distance (26), or even with the PPT placed downstream of the AG (27). Furthermore, the mammalian U2-type BPS can also be located very far (>100 nt) from the intron–exon junction sequence (28). U12-type introns also lack an obvious PPT at the 3′ss.

Historically, splice sites are ranked, based on compilations of splice sites (19,20,29,30). However, none of these ranking schemes accurately identify the bona fide splice sites (31–33). In addition, alternative splicing, involving the choice of competing splice sites, is not amenable to an analysis based solely on splice sites (34).

In order to identify common and distinguishing features in each splice-site type, we have collected and analyzed a comprehensive set of naturally-occurring splice sites from the genomes of five model organisms: *Homo sapiens*, *Mus musculus*, *D.melanogaster, Caenorhabditis elegans* and *Arabidopsis thaliana*. Since a well-classified dataset did not exist previously, we have carefully classified the splice sites into various categories, generating a user-friendly resource, SpliceRack (http://katahdin.cshl.edu:9331/SpliceRack/). Using this large-scale splice-site database, we have revisited relevant themes in splicing, such as the frequencies of U12-type and other rare introns, and the variations in the 5′ and 3′ss motifs among the five organisms. We have also applied phylogenetic approaches to infer the evolutionary histories of the different splice-site subtypes. Finally, we have developed tools, available on SpliceRack, to facilitate various splice-site analyses.

MATERIALS AND METHODS

Data collection

The idea driving the collection of data was to generate a large, reliable dataset that would allow statistical analyses. For this reason, sequences from RefSeq (35) were used instead of ESTs and cDNAs and other sources of noisy data. The RefSeq collection contains sequences with different levels of confidence (including categories, such as reviewed, provisional, predicted etc.).

RefSeq sequence data from NCBI (from June 2005) for five reference genomes was collected, including the primate *H.sapiens*, the rodent *M.musculus*, the arthropode *D.melanogaster*, the nematode *C.elegans* and the land plant *A.thaliana*. If a splice site occurred more than once in different alternatively spliced isoforms, it was considered only once in the database, as part of one of the variants, in order to avoid counting splice sites multiple times.

The flanking sequences at both exon–intron junctions were extracted, 38 nt long at the 5′ss (8 nt within the exon and 30 nt within the intron), and 43 nt long at the 3′ss (35 nt within the intron and 8 nt within the exon). The splice sites were identified using annotations of the genome from GenBank. *C.elegans* 3′ss that are *trans*-spliced to spliced leader (SL) RNAs (36) were not included in the analysis. The data, along with lengths of introns and exons, are stored in a relational database.

Position weight matrices (PWMs), information content and phylogeny

Sequences at the 5′ and 3′ junctions for each intron were separately aligned using the intron–exon junction as the anchor. The alignment was used to calculate the frequencies of nucleotides at each position. This results in a PWM that has four rows (one for each of A, C, G and T) and the number of columns is equal to the length of the splice-site motif. We chose a 9 nt long motif for the 5′ss (3 in the exon and 6 in the intron) and a 14 nt long motif for the 3′ss (13 in the intron and 1 in the exon). The consensus sequence for a PWM is constructed by choosing the nucleotide with the highest frequency of occurrence at each position.

A log-odds (LOD) matrix was constructed from the PWM by the usual method of taking a logarithm to base 2 of the ratio of frequencies at each position of the matrix against a background frequency of 0.25 for each nucleotide. Using a different background will not change the results, but will change the scores and thresholds. A position on a sequence was scored by aligning it with the matrix and adding up the entries in the matrix that correspond to the nucleotides at each position on the sequence.

Each position on the splice site has an associated information content that can be derived from the frequencies of the 4 nt, using information theory (37,38). For a random variable that can take on four values with four probabilities, p_i, the information content is $-\Sigma p_i \lg p_i$. Information content can be used to study the variability (or level of conservation) at each position, varying from 0 (one possibility is certain) to 2 (all possibilities are likely equal). Some papers use $2 + \Sigma p_i \lg p_i$ as a measure of conservation, which just inverts the scale (39).

A distance, d, can be derived between two PWMs, p and q, of equal length, using a measure called the Kullback–Leibler distance (40), a symmetrized form of which is defined as

$$d = -\Sigma_i [p_i \lg(p_i/q_i) + q_i \lg(q_i/p_i)], \quad \mathbf{1}$$

where i is the index of position on the PWM. This distance was used to derive phylogenetic trees for the 5′ and 3′ss separately, using the program Phylip (41).

Scoring a putative splice site

Given a PWM of a given splice-site type, a log-odds scoring matrix was constructed to score the splice sites. A pseudocount of 0.0001 was added to avoid logarithms of zero. The small pseudocount ensures a strong penalty for mismatches at the exon–intron junctions.

The scores obtained by each LOD score matrix were rescaled to fall in the 0–100 range. All positive scores, from 0 to the maximum, were rescaled to lie between 50 and 100, and all negative scores, from the minimum score to 0, were rescaled to lie between 0 and 50. A score close to 50 means the background model is favored, a score close to 100 favors the splice-site motif in question, and a score close to 0 implies that the splice site is different from both the background and the splice-site distribution. Finally, strongly conserved splice sites have a bigger spread of scores, so a single mismatch can drop the score dramatically, whereas less conserved splice sites are more tolerant of mismatches compared to the highest scoring sequences.

Classification of introns

PWMs were first generated for each splice-site type and BPS using human curation. The U12-type BPS consists of an 8 nt pyrimidine-rich sequence (approximately TTTTCCTT), followed by two positions enriched in As and two positions enriched in pyrimidines (C or T). The U12-type BPS requires an A, either at positions 9 or 10 (42). In order to take this into account, we constructed two U12-type BPS PWMs, one with a highly conserved A at position 9 (U12-BPS-A9), and another with a highly conserved A at position 10 (U12-BPS-A10). A sequence was considered a good BPS if it had a score greater than 65 for either one of the two BPS PWMs and if it was located between 8 and 30 nt upstream of the 3′ss.

The short fragments around the exon–intron boundaries were scored using the LOD matrices, and each intron received a score for 5′ss, 3′ss and BPS for U12-type introns. The best examples of each type, those with high 5′ss scores and consistent dinucleotide boundaries, were selected. For U12-type GT–AG introns, an additional criterion for the presence of a BPS between 8 and 38 nt upstream of the 3′ end was used to eliminate possible contaminants. The best set of introns for each subtype was used to generate new species and splice-site type specific PWMs, which were then used to rescore the splice sites and reclassify them.

Intron classification was unambiguous in almost all cases, either due to their distinctive boundaries or their 5′ss scores, with the exception of a few introns with similar scores for U2- and U12-type GT–AG 5′ss matrices. The introns that remained unclassified were subsequently classified using criteria described below.

Previous classification schemes (6,8,25) assumed that U12-dependent splicing required a good BPS within 10–20 nt from the 3′ end of the intron (22), but recent data has shown this is not always necessary (7,25,43). Thus, we used the presence of an optimal U12-type BPS only to discriminate between introns that could not be unambiguously classified by their 5′ss scores alone. If the U12-type 5′ss score was greater than the U2-type 5′ss score by more than 25 then it was assigned the U12-type GT–AG. If the U12-type 5′ss score was greater than the U2-type 5′ss score by more than 10 and the intron had a good U12-type BPS, then again it was assigned the U12-type GT–AG. In all other cases it was assigned the U2-type GT–AG.

The thresholds of 25 and 10 are arbitrary, but were optimized by human curators to minimize the error rate, which was estimated at <1% by manual scrutiny. The borderline cases could be classified in either group, and their unambiguous assignment will require experimental verification.

Finally, introns with a minimum 5′ss score less than 50 are kept in the database, and can be accessed from the SpliceRack website, but are not used in our analysis.

Introns with non-canonical boundaries

A few of the introns with dinucleotide boundaries that do not fit into the canonical categories are real, and we delineate the strategy to identify them. In general, current datasets are biased against non-canonical splice sites, but it is possible to gather some information on them from the RefSeq sequences, as we show below.

To identify non-canonical U12-type introns from this unclassified set, we chose those with good matches to the U12-type 5′ss, with a non-canonical dinucleotide at the intron's 3′ end, and a significant number of these introns were found (Table 1). Many of these introns have a U12-BPS. Analogous to the scheme outlined above, those introns without a match to the U12-type BPS were only classified as U12-type when the score for the U12-type 5′ss was >25 score units higher than that for the U2-type 5′ss. These introns were included in the U12-type RT–RN category (where R is a purine, and N is any nucleotide).

Some GC–AG introns revealed a strong match to the PWM for U12-type GT–AG 5′ss, except for the mismatch at position +2. We searched for GC–AG introns with the GCATCC motif from positions +1 to +6 of the 5′ss, and found that most of these introns had a match to the BPS at an optimal distance from the 3′ss. These introns were named U12-type GC–AG.

U2-type AT–AC introns (or AT–AC type II introns) were identified by searching for unclassified AT–AC introns that bear a good match to the U2-type 5′ss, except for the intron's terminal dinucleotides. A PWM for U2-type AT–AC 5′ss was generated from the PWM for U2-type GT–AG 5′ ss, by changing the full conservation from G to A at position +1. Only those introns with a minimal 5′ss score of 75 were reclassified as U2-type AT–AC.

Putative examples of other classes of non-canonical U2-type introns were identified by searching for all exon–intron boundaries with good 5′ GT–AG U2-type splice sites (score greater than 70), with a non-canonical dinucleotide boundary at the 3′ terminus of the intron and a good PPT near the 3′ terminus in the intron.

Introns that could not be classified into any of the seven categories either lack a canonical exon–intron boundary dinucleotide, and/or have all 5′ss scores below 40. These can be accessed through SpliceRack. Most of these unknown splice sites are probably due to sequencing or annotation errors in

Table 1. A synopsis of the collection of splice sites in SpliceRack

	H.sapiens	*M.musculus*	*D.melanogaster*	*C.elegans*	*A.thaliana*	Total
Number of accessions[a]	23 788	21 824	11 763	19 278	22 434	99 087
Number of genes[a]	23 505	21 772	11 756	19 269	22 354	98 656
U2-type GT–AG	183 678	174 671	40 637	93 699	111 351	604 036
U2-type GC–AG	1602	1439	185	351	903	4480
U2-type II (AT–AC)	15	6	3	0	0	24
Total U2-type	185 295	176 116	40 825	94 072	112 254	608 562
U12-type GT–AG	469	444	11	0*	162	1086*
U12-type AT–AC	169	151	6	0	27	353
Other U12-type[b]	33	30	1	0	2	66
Total U12-type	671	625	18	0	191	1505*
Total number	185 966	176 741	40 843	94 072	112 445	610 067
% U2-type GC–AG	0.865	0.817	0.453	0.373	0.804	0.736
% U2-type AT–AC	0.008	0.003	0.007	0	0	0.003
% U12-type	0.361	0.354	0.044	0	0.170	0.247

This table shows the total number of introns in SpliceRack for each subtype.
The * indicates that there are 22 U12-type-like GT–AG introns in *C.elegans*, which were not included in the counts of U12-type introns (see text for details).
[a]The number of accessions (RefSeqs) includes alternatively spliced isoforms.
[b]Other U12-type introns refers to those U12-type introns lacking the canonical U12-type terminal dinucleotide pairs, which were sorted in the RT–RN and GC–AG categories (see text for details).

the genome (44), as shown in previous examinations of non-canonical splice sites (45–47). The unclassified introns in our dataset represent a small fraction (<1.5%) of the total number of introns.

RESULTS AND DISCUSSION

SpliceRack

We have created the largest compilations of splice-site sequences to-date. The splice-site sequences for the five species were extracted from the RefSeq collection of NCBI (35). Introns were classified into various splice-site subtypes, based on scoring methods derived from PWMs, as explained in the Materials and Methods section (Table 1). Our data is publicly available through the SpliceRack website (see http://katahdin.cshl.edu:9331/SpliceRack/).

Tools to explore SpliceRack

SpliceRack contains sequences flanking the 5′ss (38 nt long, 8 nt within the exon and 30 nt within the intron), and the 3′ss (43 nt long, 35 nt within the intron and 8 nt within the exon). We have developed several useful tools, available on the SpliceRack website, that allow exploration of the splice sites and their flanking sequences. Each tool results in the selection of a subset of splice sites. The genes containing these splice sites can be analysed for functional and other biases by using GObar (48), which is provided as a built-in tool.

Retrieve splice sites by type. This tool allows users to retrieve splice site flanking sequences by specifying one or more splice-site subtypes, cut-off 5′ss or 3′ss scores, and species (Figure 2). A similar display is obtained in the 'Search for genes' tool, which allows visualization of all splice sites within a queried gene. Introns with weak splice-site boundaries can be retrieved using a minimum 5′ss score of 40 and a maximum of 50.

Search for patterns. This tool can be used to locate potential protein-binding sites by searching for sequences that match user-specified patterns in the form of regular expressions. For example, IUPAC SNP codes as well as standard symbols, such as '+' and '*' can be used. The search can be restricted by species, splice-site type and start position, as explained on the website.

The neural-specific splicing factor Fox-1 binds to the hexamer UGCAUG. In order to find introns with the hexamer in the vicinity downstream of the 5′ss, the database was queried for U2-type GT–AG 5′ss with the pattern N{13,31}TGCATG. The N allows match to any nucleotide, the 13 and 31 ensure that the TGCATG motif will start anywhere from position 13 on the flanking sequence, which is after the U1 binding site, to position 31. This query results in 1221 introns. By using the GObar function, we confirmed that the genes containing these motifs are enriched for function in neural tissues (49). This analysis will not find genes that have this motif deep inside the intron.

Search for motifs. This tool is similar to the search for patterns tool, except the patterns are now of fixed-length, which speeds up the search. For example, to find the frequency of potential stop codons (TAG,TTA,TAA) at the last 3 nt of an exon, the search is carried out for human U2-type GT–AG 5′ ss with a motif start position of 5. A query with the sequence TAG returns 4591 introns, while TRA (R is a purine), representing the other two stop codons, returns an additional 1439 introns.

Search for matrices. This tools finds matches to a user-supplied PWM in the sequences in SpliceRack. Users can also supply cut-off scores, and the tool calculates the frequency of the motif's occurrence, starting from a user-specified position relative to the exon–intron boundary. It also displays the rank (based on counts) of a searched motif relative to the frequencies for all the motifs that are of the length of the PWM and start at the user-specified start site. Results can be limited by species, 5′ss or 3′ss and splice-site type. This tool can be used to locate sequences in the neighborhood of splice sites that can serve as protein-binding sites.

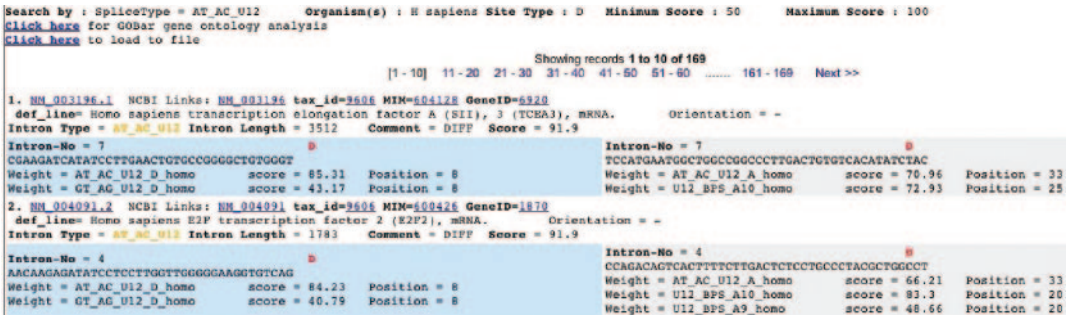

Figure 2. Two of the 169 human U12-type AT–AC introns are shown. These were accessed using a minimum 5′ ss score of 50 and a maximum of 100. The output includes links to the GenBank accession, gene names and the intron number and length. The two highest scores for each splice-site subtype are shown. The classification of introns is explained in the Materials and Methods section.

Splice-site locator. This tool can help identify potential splice sites in user-specified genes or sequences, using scores derived from the PWMs in SpliceRack. When a gene name (or sequence) is supplied along with a user-specified score threshold, the known exon–intron boundaries as well as predicted splice sites are shown. In addition, a distribution of counts of predicted splice sites against scores is also shown, so that it is easy to see how the genuine splice sites compare to the predicted splice sites. A link to the sequence in the gene where the potential splice sites are located is also provided.

U2-type splice sites

U2-type GT–AG and GC–AG splice sites. Of all the splice sites we have classified, 99.6% (608 562) are of the U2-type in the five genomes (Table 1). GC–AG introns represent 0.82% of U2-type introns in *H.sapiens*, 0.86% in *M.musculus*, 0.8% in *A.thaliana*, 0.45% in *D.melanogaster* and 0.37% in *C.elegans*, indicating that more primitive organisms tend to have fewer GC–AG introns.

GT–AG and GC–AG subtypes of U2-type introns are considered separately. Though there is no evidence that these types are spliced by distinct mechanisms, their PWMs are significantly different (Figure 3). The 5′ss from GC–AG and GT–AG introns bind to U1 snRNP during spliceosome assembly (2). The consensus sequence for the U2-type GT–AG 5′ss motif corresponds to a perfect base pairing to the U1 snRNA 5′ end. Despite this, a high degree of variability seems to be tolerated at most positions of the actual 5′ss. In GC–AG introns, the substitution at position +2 of the 5′ ss introduces a mismatch in the U1: 5′ss helix. The PWMs (Figure 3) show that the remainder of the 5′ss sequence has a higher degree of match to the consensus GT–AG 5′ss, apparently compensating for the mismatch at position +2 (17,23,24,50,51). Subtype switching of introns from U2-type GT–AG to U2-type GC–AG between human, rodent and chicken has been shown to be a common event (24).

Non-canonical U2-type introns. Non-canonical U2-type introns with AT–AC dinucleotide intron ends have been known to occur and are functional.

This is the first genome-wide survey for U2-dependent AT–AC II introns (Table 1, Materials and Methods). Previously, only two human AT–AC introns in the *SCN4A* and *SCN5A* genes were experimentally demonstrated to be spliced by the U2- but not the U12-type splicing machinery (5,51). The borders of these introns were conserved between human and puffer fish. In addition, the chicken parvalbumin gene and some xylanase genes in filamentous fungi were hypothesized to contain AT–AC II introns. Furthermore, a G to A transition at the first intronic nucleotide can be rescued for splicing by a G to C transversion at the last nucleotide of U2-type introns in yeast (52,53) and human pre-mRNAs (54), although less efficiently.

We found a total of 15 AT–AC II introns in the human, 6 in the mouse and 3 in the *D.melanogaster* genomes. The case for these introns is bolstered by the fact that they are preserved in orthologous genes, such as the AT–AC II intron in the human and *D.melanogaster TAF2* gene. In the human dataset, 10 out of 15 of these introns are found in members of the alpha subunit of the voltage-gated sodium channel gene family, which also have U12-type introns. Hence, the higher number of AT–AC II introns in mammals could be explained by the expansion of this gene family. All six mouse AT–AC II introns are found in members of this family, and conserved in the human genome. Interestingly, AT–AC II introns seem to be absent in *C.elegans* and *A.thaliana*. The very low abundance of these introns, and their presence in vertebrates and fungi, suggest that they constitute a relic of U2-dependent introns that have been progressively lost.

We searched for other putative non-canonical U2-type splice sites as explained in the Materials and Methods section. Many of the known types (55) (see http://www.tigr.org/tdb/e2k1/ath1/Arabidopsis_nonconsensus_splice_sites.shtml) are not in our compilation as they do not occur in the RefSeq collection. We found 99 introns (33 in human, 46 in mouse, 5 in fly, 11 in the worm and 4 in the plant) with an optimal U2-type 5′ss and PPT, but a non-canonical 3′ intron terminus. Out of these, the GT–AT and GT–TG introns in the DGCR2 and ARS2 genes, respectively, are conserved between human and mouse, suggesting that at least some of the remaining 95 might be bonafide, functional splice sites.

U12-type splice sites

Canonical U12-type introns. We have substantially expanded the catalogue of U12-type splice sites in four of the five organisms we examined (Table 1). The PWMs for these splice sites have been derived and are consistent with previous reports (6,25) (Figure 4). However, we find a more

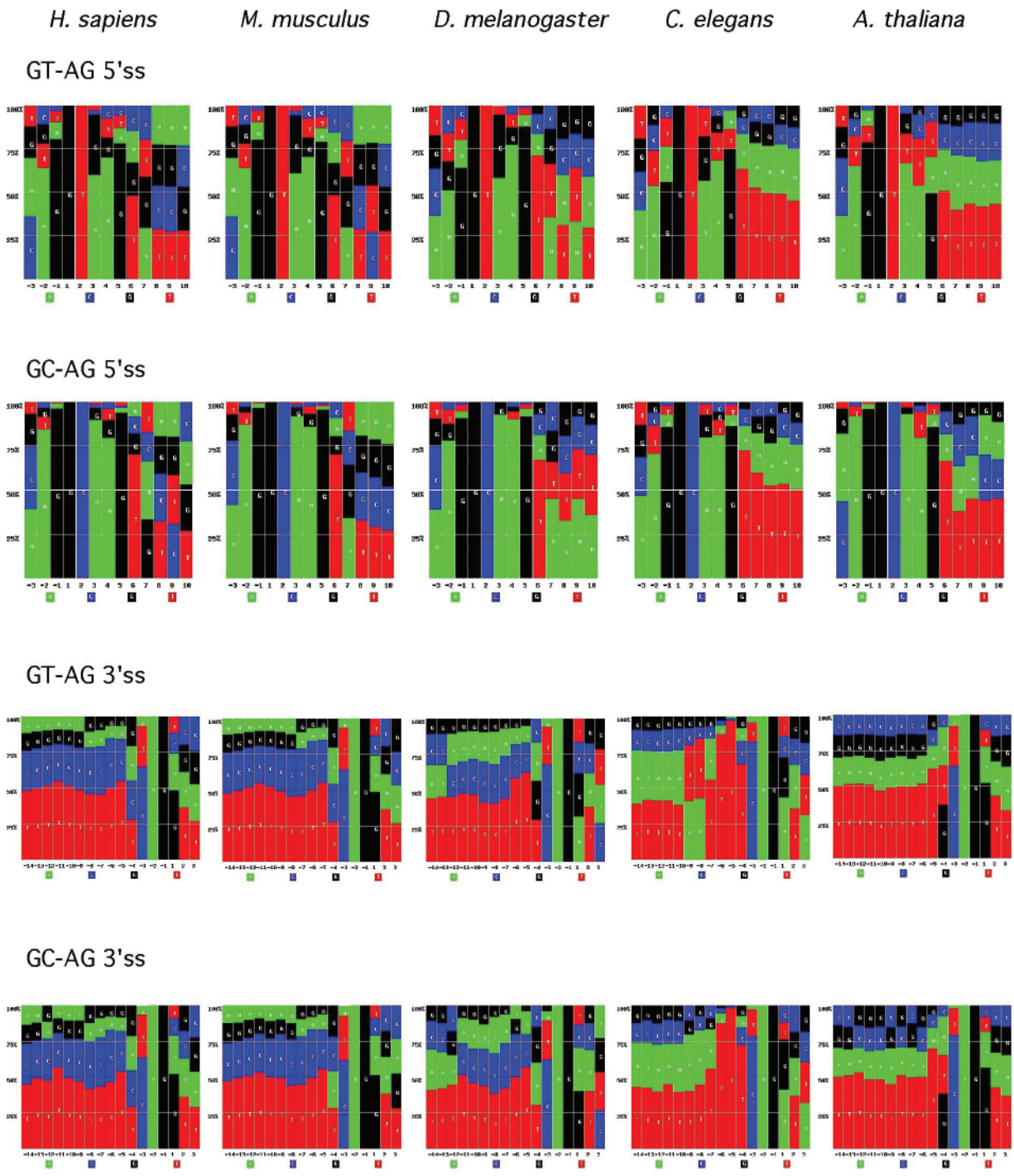

Figure 3. PWMs for 5′ and 3′ ss for the U2-type GT–AG and GC–AG intron subtypes, for the five species. We used color-coded vertical bars stacked one over the other to represent the percentage of each nucleotide, Green (A), Blue (C), Black (G) and Red (T); similar logos have been used elsewhere (66). The bars are ordered by their percentages.

extended region of conservation than previously identified (6,22,25). Specifically, the −1 position on 5′ss has a strong preference for T in U12-type GT–AG, and the three exonic positions immediately downstream of the 3′ss favor the sequence ATA.

We have identified 671 human U12-type introns, ∼0.36% of all introns, which includes 314 introns that were hitherto unrecognized as U12-type introns. The GT–AG subtype accounts for 469 introns while 169 introns belong to the AT–AC subtype, and the remaining 33 lack the canonical U12-type intron borders (Table 1). We have identified 357 of the 404 U12-type introns predicted in the most recent previous survey of U12-type introns (6) while the remainder are from genes not present in the RefSeq collection. Mouse frequencies for U12-type introns are very similar to those for human, consistent with a previous study (24).

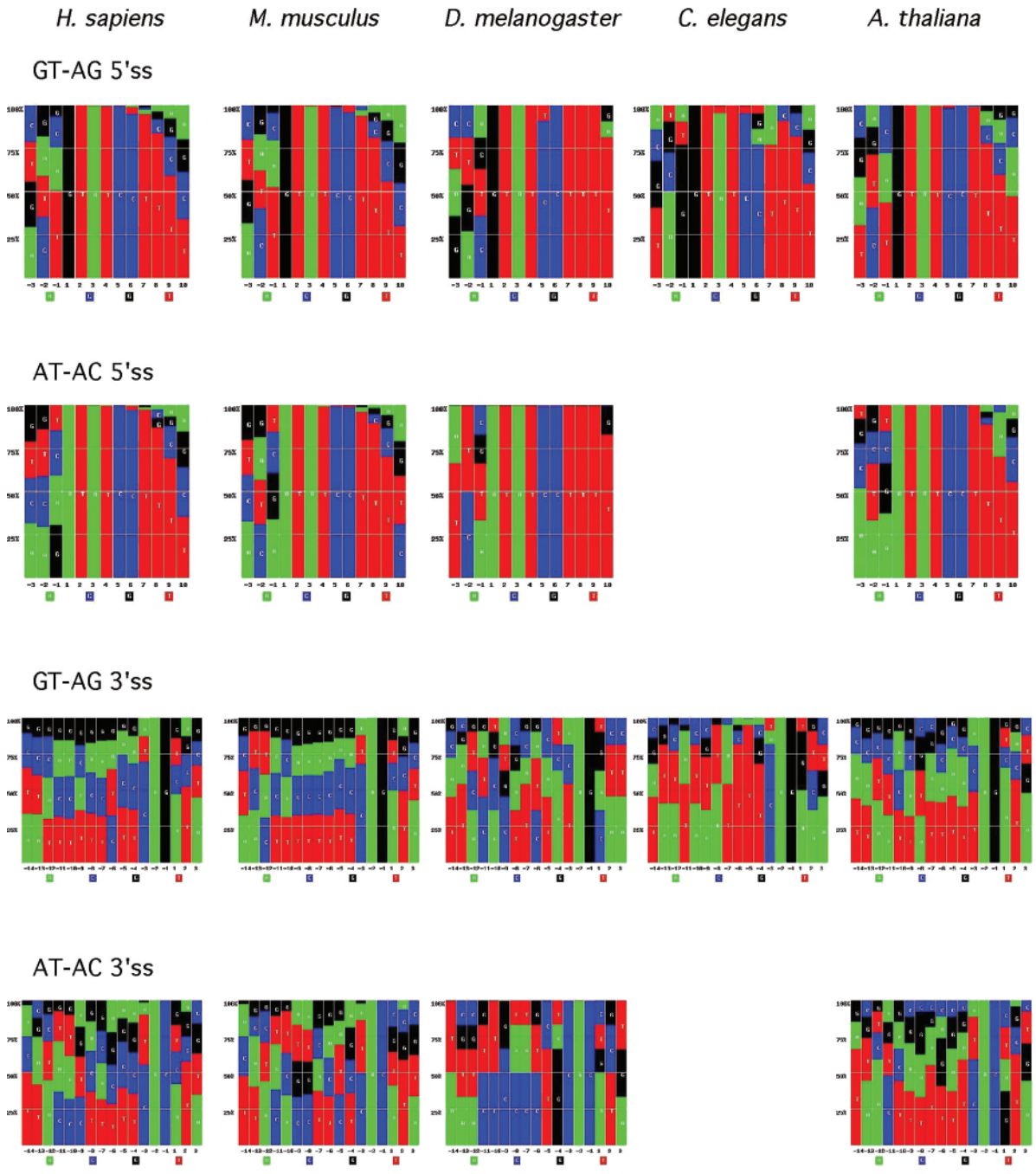

Figure 4. PWMs for 5′ and 3′ ss for the GT–AG and AT–AC U12-type intron subtypes, in the five species. For *C.elegans* what is shown is the vestigial U12-type GT–AG splice sites.

We identified 189 U12-type introns in *A.thaliana* (Table 1), as compared to the 165 U12-type introns derived earlier from a mapping of cDNA/ESTs to the genomic sequence (25). Overall, the fraction (0.24%) of U12-type introns in *A.thaliana* is smaller than the fraction in mammals.

We found 18 U12-type introns in *D.melanogaster* (Table 1), of which 6 are of AT–AC subtype, 11 are of GT–AG subtype and 1 has GC–AG terminal dinucleotide (see below). A previous study identified 19 U12-type introns in *D.melanogaster* (56). We are missing one of their AT–AC introns and three of their GT–AG introns, either due to the genes not being in RefSeq or due to the sequences not having a good BPS. Even though U12-type introns in *D.melanogaster* correspond to a very small fraction (0.04%) of all introns, in at least some cases they are essential for the development of this organism (57). Strikingly, the relative abundance of each U12 subtype in the fly is very different from that of mammals and plants, in that *D.melanogaster* has a greater percentage of AT–AC U12-type introns compared to GT–AG U12-type introns than the other three species (33 versus 14–25%).

U12-type introns have not been found in *C.elegans* (8) and recent systematic searches for U12-type splicing factors

failed to identify proteins unique to the U12-type spliceosome in the *C.elegans* genome (1,58). Since other metazoans and plants have U12-type introns, it has been suggested that the *C.elegans* lineage lost the U12-type introns (8). Indeed, we found some vestigial U12-type *C.elegans* introns whose 5′ss are strongly reminiscent of U12-type GT–AG introns in other species (Table 1).However, because the 3′ end sequences of these introns have the signature features of *C.elegans* U2-type 3′ss (see below), and a good candidate sequence for a U12-type BPS is missing, they are probably spliced by the U2-type machinery.

Non-canonical U12-type introns. A significant fraction of U12-type introns in our dataset (≈5%) lack the canonical pairs of dinucleotide (Table 2, see Materials and Methods). These transcripts are supported by RefSeq sequences. However, a large fraction of these introns are incorrectly mapped in other genome browsers, probably due to automated mapping based on canonical splice-site predictions.

Only two of these introns were found in *A.thaliana*, including the previously described intron 7 of the AtG5 gene (59). Such non-canonical introns have been experimentally shown to be competent for splicing by the U12-type machinery (7). The most efficient pairs were found to be GT–AG and AT–AC, but AT–AN (where N is any nucleotide) and GT–AT could also engage in productive splicing. Remarkably, the splicing efficiency of these splice-site combinations correlates with their frequencies in our database. Furthermore, evidence has been found for U12-type non-canonical splicing in samples from a Peutz–Jeghers syndrome patient that have a A to G mutation at the +1 position of the AT–AC U12-type 5′ ss in the *LKB1* gene (43).

We have also identified another subtype of putative U12-type introns with GC–AG borders, which were predicted in an earlier survey (8). There are four candidates of these introns in the human and mouse, and one in the fly. All of these candidates have a perfectly conserved GCATCCT 5′ ss motif, and most have an optimal U12-type BPS at an appropriate location. Furthermore, the exonic nucleotide upstream of the 5′ss of most of these candidates do not usually conform to the U2-type consensus making them poor substrates for U2-type splicing.

Biases in the distribution of U12-type and other rare introns. The persistence of U12-type introns over evolutionary time is a mystery, as they represent a very small fraction of all genes, and some lineages and organisms, such as *C.elegans*, appear to have completely lost them. We used GObar (48) (see http://katahdin.cshl.org:9331/GO/GO.cgi), a tool that exploits the information in Gene Ontology to unveil pathways or functions that are enriched in a particular sample gene set, to study the functional distribution of genes containing U12-type introns. We find that U12-type introns are usually found in genes involved in ion transport (25,60), protein trafficking and cell-cycle control, among other processes.

Genes with more than one U12-type intron are found more often than expected from a random model for the disribution of U12-type introns in genes, consistent with previous reports (6,8). *H.sapiens* and *M.musculus* have 45 and 46 genes that contain two U12-type introns, respectively, *A.thaliana* has only six, and *D.melanogaster* has none. Furthermore, human and mouse have two and three genes, respectively, with three U12-type introns, all of them members of the solute carrier family (SLC) gene family, that encode sodium and hydrogen exchanger proteins. In mammals, considering an average number of eight introns per gene (61) and assuming an independent U12-intron distribution model, the expected number of genes with two U12-type introns is only six. If we include more realistic distributions of introns amongst genes, the total expected goes up to only 16.

This physical clustering of U12-type introns could be related to function or to the origin of U12-type introns. The U12-type machinery is ≈100 times less abundant in the nucleus than the U2-type machinery (62,63) and a possible consequence of this is slower splicing kinetics of U12-type introns. Indeed, there is some experimental corroboration for this idea (64). These observations suggest that U12-type introns persist because of their slower rates of splicing reaction, which might be important for the expression of the gene that contains them. If the U12-type intron is the slowest step in the pre-mRNA processing, having more than one U12-type intron in a gene might have no additional impact on splicing dynamics.

The possibility of intron conversion from U12- to U2-type was previously demonstrated by looking at orthologous U12-type introns from different species (8,60), but the rate of conversion is unknown. As pointed out in these studies, the high conservation of U12-type splice sites and the high degeneracy of U2-type splice sites suggests that this conversion could be achieved by a few mutations, and that the reciprocal conversion would be very difficult. Interestingly, we noticed frequent subtype switching of non-canonical U12-type introns between human and mouse. Out of 24 conserved U12-type introns between human and mouse of which at least one of the two is non-canonical, only 10 showed the same intron borders (Table 3) The remaining 14 introns underwent subtype switching. Amongst these, we found two cases of subtype switching from a non-canonical intron pair to another one, and ten introns that were switched from a non-canonical intron pair to a canonical GT–AG or AT–AC intron or vice versa. Remarkably, all these non-conserved intron borders differ at only one position.

The AT–AG U12-type introns might represent viable intermediates of subtype switching events between U12-type AT–AC and GT–AG introns. In light of this, the >50% subtype switching for the non-canonical introns seem to be at odds with the observation that no subtype switching between

Table 2. The distribution of non-canonical U12-type introns

	H.sapiens	*M.musculus*	*D.melanogaster*	*A.thaliana*	Total
AT–AG	11	8	0	1	20
AT–AA	3	4	0	1	8
AT–AT	6	6	0	0	12
AT–GT	0	1	0	0	1
GT–AT	4	3	0	0	7
GT–GG	5	3	0	0	8
GT–TG	0	1	0	0	1
GC–AG	4	4	1	0	9
Total	33	30	1	2	66
% of all U12-type	4.92	4.80	5.56	1.05	

Table 3. Subtype switching for non-canonical U12-type introns between human and mouse

	H.sapiens	*M.musculus*	Subtype
Human gene			
XPO4	AT–AG	GT–AG	Non-canonical/canonical
ACTR10	AT–AG	GT–AG	Non-canonical/canonical
NUP210	AT–AG	AT–AG	Conserved
SCN10A	AT–AG	GT–AG	Non-canonical/canonical
CUL4B	AT–AG	AT–AT	Non-canonical/non-canonical
USP21	AT–AG	AT–AG	Conserved
RAPGEFL1	AT–AG	AT–AG	Conserved
INPP5B	AT–AG	AT–AG	Conserved
RASGRP4	AT–AA	AT–AA	Conserved
FLJ21106	AT–AA	AT–AA	Conserved
MSH3	AT–AA	AT–AC	Non-canonical/canonical
ERCC5	AT–AT	AT–AC	Non-canonical/canonical
IK	AT–AT	AT–AA	Non-canonical/non-canonical
DEADC1	GT–AT	GT–AT	Conserved
SLC24A6	GT–AT	GT–AT	Conserved
SLC12A7	GT–GG	GT–AG	Non-canonical/canonical
ARAF	GT–GG	GT–GG	Conserved
LZTR1	GC–AG	GC–AG	Conserved
Mouse gene			
4930449E07Rik	**AT–AC**	AT–AG	Non-canonical/canonical
Cacna1h	**AT–AC**	AT–AG	Non-canonical/canonical
4930590J08Rik	**GT–AG**	AT–AG	Non-canonical/canonical
Gpaa1	**AT–AC**	AT–AT	Non-canonical/canonical
Zc3hdc6	**AT–AC**	AT–AT	Non-canonical/canonical
Slc9a8	**AT–AC**	AT–AT	Non-canonical/canonical

The canonical borders, AT–AC or GT–AG, for one of the species are shown in bold letters. For non-conserved introns, the differences in the nucleotides on the intron borders are underlined. There are 10 cases where the non-canonical boundaries are conserved between mouse and human, and 2 cases where one non-canonical form switched to another non-canonical form between the species. In the rest of the cases, one of the boundaries, either the mouse or human is a canonical form.

U12-type AT–AC and GT–AG intron orthologs was found between human and chicken (24)

Splice-site analysis by comparative genomics

We used SpliceRack to analyze the species-specific properties of splice sites from an evolutionary perspective. The global patterns of the splice sites were compared amongst the five species by phylogenetic approaches.

Splice-site information content across species. We measured and plotted the information content at each position of the 5′ and 3′ss from the PWMs (Figures 5 and 6 see Materials and Methods). The information content is a measure of the degree of conservation in splice-site sequences, with lower information content corresponding to a higher degree of conservation, and vice versa (24,65–68).

Information content of U2-type 5′ss (Figure 5): The 5′ss motifs for the two mammals are virtually identical, as described previously (24,69). The U2-type GT–AG 5′ss graph ss graph shows that there is more conservation in the exonic nucleotides for *H.sapiens*, *M.musculus* and *A.thaliana*, compared to the other two simpler species, *D.melanogaster* and *C.elegans*. This is consistent with the fact that budding and fission yeast have poor conservation at the exonic 5′ss positions, but a very strict conservation at intronic positions (70). *C.elegans* and *A.thaliana* show weak conservation of Ts at positions downstream of +6, which is not the case for the other organisms (8,71). The GC–AG 5′ss show higher conservation at both intronic and exonic positions when compared to the GT–AG 5′ss (23,24,45).

Information content of U2-type 3′ss (Figure 6): The U2-type GT–AG and GC–AG 3′ss show similar information content, suggesting they might be functionally very similar. All species, except for *C.elegans*, show little conservation in the exonic or intronic portions of the 3′ss, probably due to the degenerate PPT. *C.elegans* exhibits a shorter and more conserved PPT in the intronic portion, with a consensus sequence of TTTTCAGR, where AG is the terminus of the intron (72). The existence of this conserved 3′ss motif in *C.elegans* is explained by a more stringent requirement for an optimal binding site for the U2AF heterodimer that is unique to this species (73). The conservation of G at position +1, conserved only in *H.sapiens*, *M.musculus* and *A.thaliana*, can be seen as a small decrease in information content at that position for the three species.

Information content of U12-type 5′ss (Figure 5): The information content distribution for U12-type GT–AG and AT–AC 5′ss shows a high degree of conservation at the intronic positions that starts decaying at position +6. The conservation of T at position −1 of mammalian and plant, but not fly U12-type GT–AG 5′ss can be seen as a small increase in the fly's information content at that position. The information content distribution for *C.elegans* U12-type-like GT–AG introns is kept in the graph for comparison.

Information content of U12-type 3′ss (Figure 6): The information content distributions for both U12-type GT–AG and AT–AC 3′ss show a very small degree of conservation in the intronic nucleotides. This observation is consistent with the notion that U12-type 3′ss are slightly enriched in pyrimidines but lack a PPT. The location of a BPS near (8–14 nt) the 3′ end of the U12-type intron contributes to conservation at these positions, as seen in the slight decrease of information content on the left of these graphs. The small numbers of these types of introns makes it difficult to infer general trends for *D.melanogaster* and *A.thaliana*'s AT–AC 3′ss. The vestigial U12-type-like introns in the nematode also contain the conserved motif of the U2-type 3′ss, consistent with the view that these introns are indeed U2-dependent.

Phylogenetic trees from PWMs

The Kullback–Leibler distance (40) (see Materials and Methods) between PWMs for 5′ and 3′ss was used to calculate a distance matrix between species, and phylogenetic trees were derived from these distance matrices. It is possible to use other methods to build trees, but this is the simplest one (50). Trees were generated for all the splice sites together (Figure 7), after removing the conserved dinucleotide terminus of the introns from consideration, as they would hide the signal from the other positions on the splice-sites.

From Figure 7, it is easy to see a clear separation between U2- and U12-type 5′ss motifs. The various 5′ splice types separate according to splice-site type, rather than species. All the U2-type motifs cluster together, with a small separation between GC–AG and GT–AG types suggesting that the GC–AG subtype is a specialized version of the U2-type

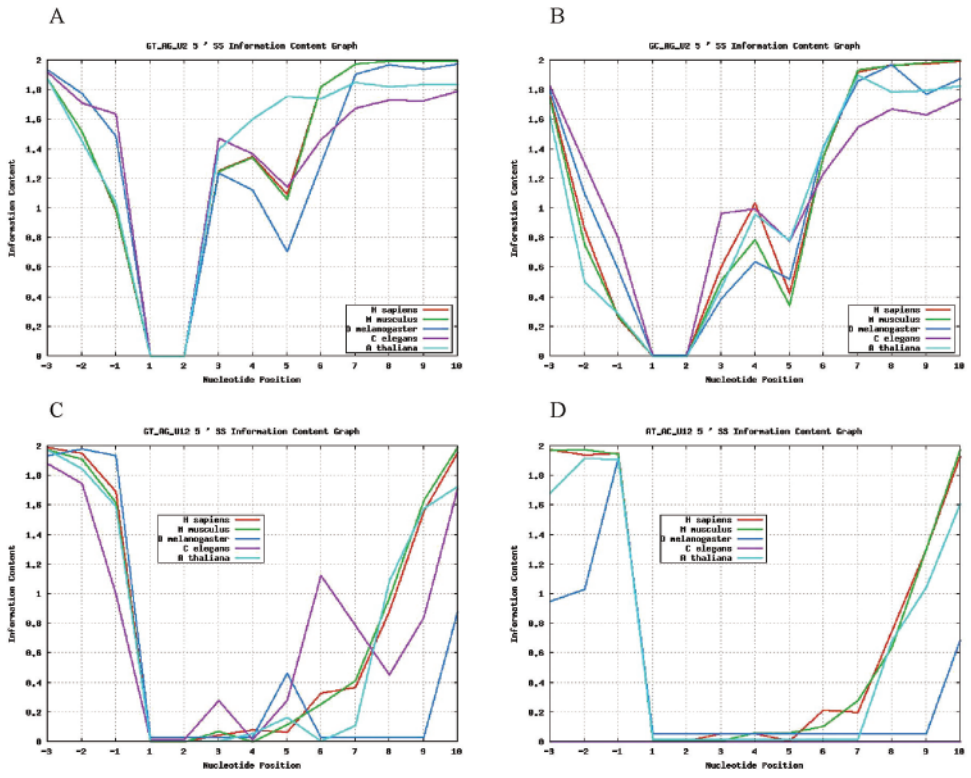

Figure 5. Information content in the exonic and intronic portions of 5′ ss in various organisms. (**A**) U2-type GT–AG introns. (**B**) U2-type GC–AG introns. (**C**) U12-type GT–AG introns. (**D**) U12-type AT–AC introns.

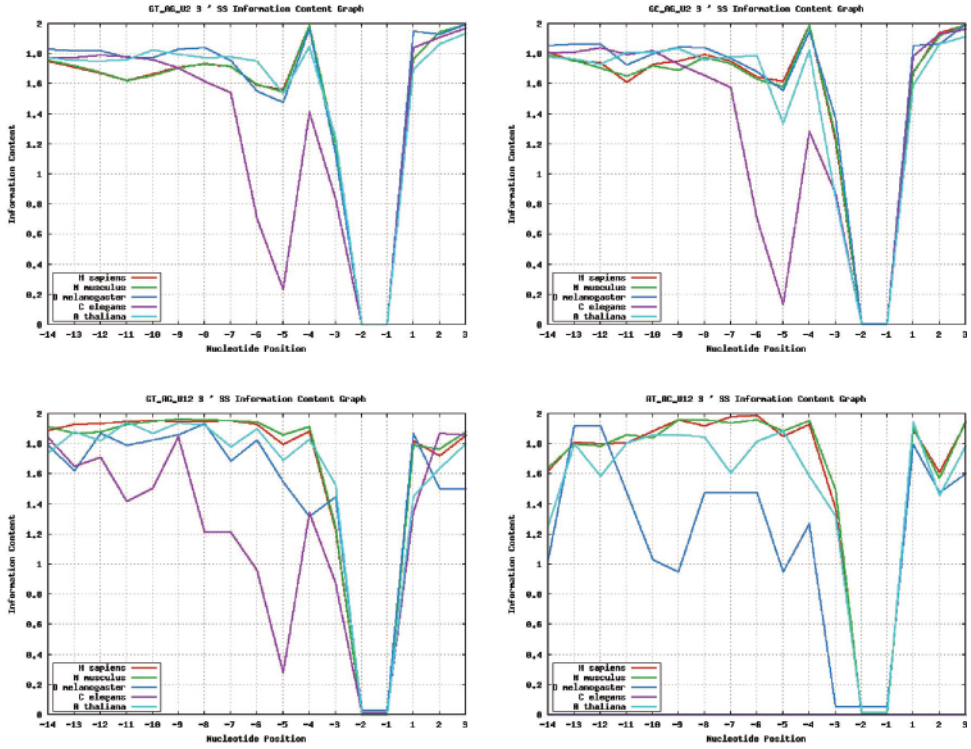

Figure 6. Information content in the exonic and intronic portions of 3′ ss in various organisms. (**A**) U2-type GT–AG introns. (**B**) U2-type GC–AG introns. (**C**) U12-type GT–AG introns. (**D**) U12-type AT–AC introns.

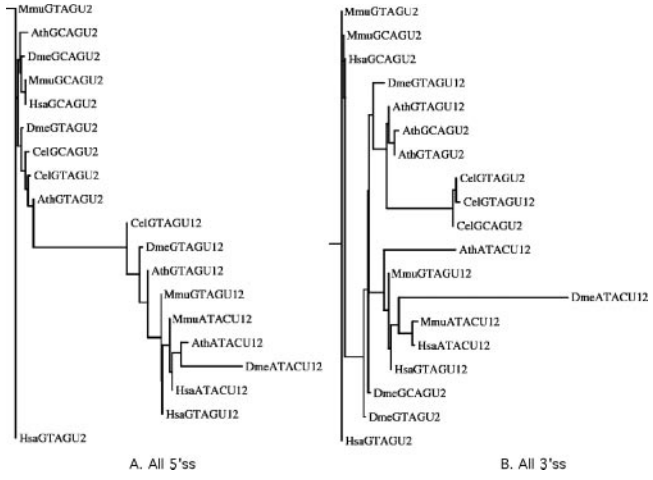

Figure 7. Phylogenetic trees for all splice site types. Trees were constructed in a manner similar to the trees for the individual splice site subtypes, but the terminal dinucleotides of the intron were removed from the PWM (see text for details).

GT–AG 5′ss motif. The *C.elegans* GC–AG motif is not as conserved as the other GC–AG motifs, so it clusters closer to the GT–AG motif from the same species. The two *C.elegans* U2-type motifs and the *A.thaliana* U2-type GT–AG motif are close to each other due to the T-rich composition of plant and nematode introns.

The 5′ss motifs for U12-type introns are distant from the U2-type introns, their closest relative being the GT–AG 5′ss motif from *A.thaliana*, probably because these motifs are rich in pyrimidines from positions +7 to +10. The *C.elegans* U12-type-like GT–AG motif is located between the U2-type and the bona fide U12-type GT–AG motifs, but it is closer to the U12-type motifs in the other species, consistent with the idea that these introns could represent a small group of vestigial U12-type introns that were converted to the U2-type. Interestingly, in this tree the four AT–AC motifs are located on the same branch of the tree, because these motifs are slightly more conserved than those for U12-type GT–AG introns.

The tree for 3′ss shows that, in general, the 3′ss tend to cluster by species rather than by subtype. Most remarkably, the three subtypes of *C.elegans* 3′ss motifs, U2-type GT–AG, GC–AG and U12-type-like GT–AG, occupy a distinct branch on the tree, indicating that: (i) these motifs are strongly related to each other, in turn suggesting that the U12-type-like introns are indeed U2-dependent; (ii) these motifs are separated from the rest, in agreement with the idea of divergent evolution of *C.elegans* 3′ss. *C.elegans* exhibits two distinct features, a short and highly conserved PPT in 3′ss (72,73) and *trans*-splicing (36). The selection of 3′ss for efficiency in *cis*- as well as *trans*-splicing might explain these features and the divergent evolution. Three of the motifs in *A.thaliana*, U2-type GT–AG and GC–AG and the U12-type GT–AG, also cluster together.

In summary, the phylogenetic trees for the different splice-site motifs reflect the evolutionary relationship between the corresponding splice sites in the five species. The clustering of the 5′ss motifs by splice-site subtype strongly suggests that these subtypes were present in a common ancestor of these five eukaryotes, and have evolved separately. In contrast, 3′ss subtypes tend to cluster by species, suggesting a more recent origin for many of the 3′ss.

Since U2-type 5′ss is initially recognized by its base pairing to the U1 snRNA which is a more specific interaction, while the 3′ss is defined by a more non-specific binding of the U2AF heterodimer to the PPT and AG element, which in turn stabilizes base pairing between U2 snRNA and the BPS. The different properties of RNA–RNA interactions at the 5′ss versus RNA–protein interactions at the 3′ss imply a greater flexibility in 3′ss selection. In support of this view we note that even within a species, there is more leeway within the 3′ss for specification of the PPT, the AG and the upstream BPS (10,26–28).

CONCLUSIONS

We have created a reliable dataset of genomic splice sites from five model organisms, based on RefSeq genes and their mappings. We have classified these splice sites into various categories, based on carefully curated PWMs for each species and splice-site type. Non-canonical splice sites, that do not fit into the four major categories, get a short shrift in databases that use large-scale automated annotation, but we have considered these cases carefully. This comprehensive set of splice sites allows robust statistical analyses, while minimizing errors. We have implemented a variety of tools, which can be used to further analyze the splice sites in our collection.

Experiments can never reproduce the entire range of phenomena that occur in nature, nor can they easily detect certain subtle features that can only be revealed by statistical analysis of a large dataset. For these reasons, the genomic dataset provides an ideal ‘dry-bench’ to study global patterns in splice sites. In addition, by using the information content of the splice-site motifs and the phylogenetic distance derived from them, we draw inferences on the evolution of the splice sites, which are only possible with large and reliable datasets. We believe this dataset and its analyses will spur further experimental work.

ACKNOWLEDGEMENTS

We thank Dhruv Pant for help with GObar and phylogenetic trees, Andrew Olson for help in linking GeneSeer to SpliceRack and James Gurtowski for developing the logo program used to depict the splice-site motifs. We also thank the anonymous reviewers for suggestions that improved the paper. X.R., M.L.H. and A.R.K. acknowledge support from NIH grants CA13106 and GM42699. Funding to pay the Open Access publication charges for this article was provided by Cancer Center, CSHL.

Conflict of interest statement. None declared.

REFERENCES

1. Collins,L. and Penny,D. (2005) Complex spliceosomal organization ancestral to extant eukaryotes. *Mol. Biol. Evol.*, **22**, 1053–1066.
2. Brow,D.A. (2002) Allosteric cascade of spliceosome activation. *Annu. Rev. Genet.*, **36**, 333–360.
3. Hastings,M.L. and Krainer,A.R. (2001) Pre-mRNA splicing in the new millennium. *Curr. Opin. Cell. Biol.*, **13**, 302–309.

4. Will,C.L. and Luhrmann,R. (2001) Spliceosomal UsnRNP biogenesis, structure and function. *Curr. Opin. Cell. Biol.*, **13**, 290–301.
5. Wu,Q. and Krainer,A.R. (1997) Splicing of a divergent subclass of AT-AC introns requires the major spliceosomal snRNAs. *RNA*, **3**, 586–601.
6. Levine,A. and Durbin,R. (2001) A computational scan for u12-dependent introns in the human genome sequence. *Nucleic Acids Res.*, **29**, 4006–4013.
7. Dietrich,R.C., Fuller,J.D. and Padgett,R.A. (2005) A mutational analysis of U12-dependent splice site dinucleotides. *RNA*, **11**, 1430–1440.
8. Burge,C.B., Padgett,R.A. and Sharp,P.A. (1998) Evolutionary fates and origins of U12-type introns. *Mol. Cell*, **2**, 773–785.
9. Reed,R. (2000) Mechanisms of fidelity in pre-mRNA splicing. *Curr. Opin. Cell. Biol.*, **12**, 340–345.
10. Moore,M.J. (2000) Intron recognition comes of age. *Nature Struct. Biol.*, **7**, 14–16.
11. Staley,J.P. and Guthrie,C. (1998) Mechanical devices of the spliceosome: motors, clocks, springs, and things. *Cell*, **92**, 315–326.
12. Patel,A.A. and Steitz,J.A. (2003) Splicing double: insights from the second spliceosome. *Nature. Mol. Cell. Biol.*, **4**, 960–970.
13. Wassarman,K.M. and Steitz,J.A. (1992) The low-abundance u11 and u12 small nuclear ribonucleoproteins (snRNPs) interact to form a two-snRNP complex. *Genome Res.*, **12**, 1276–1285.
14. Frilander,M.J. and Steitz,J.A. (1999) Initial recognition of U12-dependent introns requires both U11/5′ splice-site and U12/branchpoint interactions. *Genes Dev.*, **13**, 851–863.
15. Tarn,W.Y. and Steitz,J.A. (1997) Pre-mRNA splicing: the discovery of a new spliceosome doubles the challenge. *Trends Biochem. Sci.*, **22**, 132–137.
16. Frilander,M.J. and Steitz,J.A. (2001) Dynamic exchanges of RNA interactions leading to catalytic core formation in the U12-dependent spliceosome. *Mol. Cell*, **7**, 217–226.
17. Mount,S.M. (2000) Genomic sequence, splicing and gene annotation. *Am. J. Hum. Genet.*, **67**, 788–792.
18. Mount,S.M. (1982) A catalogue of splice junction sequences. *Nucleic Acids Res.*, **10**, 459–472.
19. Shapiro,M.B. and Senapathy,P. (1987) RNA splice junctions of different classes of eukaryotes: sequence statistics and functional implications in gene expression. *Nucleic Acids Res.*, **15**, 7155–7174.
20. Senapathy,P., Shapiro,M.B. and Harris,N.L. (1990) Splice junctions, branch point sites, and exons: sequence statistics, identification, and applications to genome project. *Meth. Enzymol.*, **183**, 252–278.
21. Hall,S.L. and Padgett,R.A. (1994) Conserved sequences in a class of rare eukaryotic nuclear introns with non-consensus splice sites. *J. Mol. Biol.*, **239**, 357–365.
22. Sharp,P.A. and Burge,C.B. (1997) Classification of introns: U2-type or U12-type. *Cell*, **91**, 875–879.
23. Burset,M., Seledtsov,I. and Solovyev,V. (2001) Splicedb: database of canonical and non-canonical mammalian splice sites. *Nucleic Acids Res.*, **29**, 255–259.
24. Abril,J.F., Castelo,R. and Guigo,R. (2005) Comparison of splice sites in mammals and chicken. *Genome Res.*, **15**, 111–119.
25. Zhu,W. and Brendel,V. (2003) Identification, characterization and molecular phylogeny of U12-dependent introns in the *Arabidopsis thaliana* genome. *Nucleic Acids Res.*, **31**, 4561–4572.
26. Merendino,L., Guth,S., Bilbao,D., Martinez,C. and Valcarcel,J. (1999) Inhibition of msl-2 splicing by sex-lethal reveals interaction between U2af35 and the 3′ splice site AG. *Nature*, **402**, 838–841.
27. Lallena,M.J., Chalmers,K.J., Llamazares,S., Lamond,A.I. and Valcarcel,J. (2002) Splicing regulation at the second catalytic step by sex-lethal involves 3′ splice site recognition by spf45. *Cell*, **109**, 285–296.
28. Smith,C.W. and Nadal-Ginard,B. (1989) Mutually exclusive splicing of alpha-tropomyosin exons enforced by an unusual lariat branch point location: implications for constitutive splicing. *Cell*, **56**, 749–758.
29. Brunak,S., Engelbrecht,J. and Knudsen,S. (1991) Prediction of human mRNA donor and acceptor sites from the DNA sequence. *J. Mol. Biol.*, **220**, 49–65.
30. Yeo,G. and Burge,C.B. (2004) Maximum entropy modeling of short sequence motifs with applications to RNA splicing signals. *J. Comput. Biol.*, **11**, 377–394.
31. Roca,X., Sachidanandam,R. and Krainer,A.R. (2003) Intrinsic differences between authentic and cryptic 5′ splice sites. *Nucleic Acids Res.*, **31**, 6321–6333.
32. Roca,X., Sachidanandam,R. and Krainer,A.R. (2005) Determinants of the inherent strength of human 5′ splice sites. *RNA*, **11**, 683–698.
33. Thanaraj,T.A. (2000) Positional characterisation of false positives from computational prediction of human splice sites. *Nucleic Acids Res.*, **28**, 744–754.
34. Matlin,A.J., Clark,F. and Smith,C.W. (2005) Understanding alternative splicing: towards a cellular code. *Nature Rev. Mol. Cell. Biol.*, **6**, 386–398.
35. Pruitt,K.D., Tatusova,T. and Maglott,D.R. (2005) NCBI reference sequence (RefSeq): a curated non-redundant sequence database of genomes, transcripts and proteins. *Nucleic Acids Res.*, **33**, D501–D504.
36. Hastings,K.E. (2005) Sl trans-splicing: easy come or easy go? *Trends Genet.*, **21**, 240–247.
37. Shannon,C.E. (1948) A mathematical theory of communication. *Bell Syst. Tech. J.*, **27**, 379–423 and 623–656.
38. Yeung,R.W. (2002) A First Course in Information Theory. In Wolf,J.K. (ed.), *Information Technology: Transmission, Processing, and Storage*. Springer (Kluwer Academic/Plenum Publishers), p. 434.
39. Schneider,T.D. (1997) Sequence walkers: a graphical method to display how binding proteins interact with DNA or RNA sequences. *Nucleic Acids Res.*, **25**, 4408–4415.
40. Kullback,S. and Leibler,R.A. (1951) On information and sufficiency. *Ann. Math. Stat.*, **22**, 7986.
41. Felsenstein,J. (1989) Phylip—phylogeny inference package (version 3.2). *Cladistics*, **5**, 164–166.
42. McConnell,T., Cho,S.J., Frilander,M.J. and Steitz,J.A. (2002) Branchpoint selection in the splicing of U12-dependent introns *in vitro*. *RNA*, **8**, 579–586.
43. Hastings,M.L., Resta,N., Traum,D., Stella,A., Guanti,G. and Krainer,A.R. (2005) An LKB1 AT-AC intron mutation causes Peutz-Jeghers syndrome via splicing at noncanonical cryptic splice sites. *Nature Struct. Mol. Biol.*, **12**, 54–59.
44. Cocquet,J., Chong,A., Zhang,G. and Veitia,R.A. (2006) Reverse transcriptase template switching and false alternative transcripts. *Genomics*, **88**, 127–131.
45. Jackson,I.J. (1991) A reappraisal of non-consensus mRNA splice sites. *Nucleic Acids Res.*, **19**, 3795–3798.
46. Burset,M., Seledtsov,I.A. and Solovyev,V.V. (2000) Analysis of canonical and non-canonical splice sites in mammalian genomes. *Nucleic Acids Res.*, **28**, 4364–4375.
47. Chong,A., Zhang,G. and Bajic,V.B. (2004) Information for the coordinates of exons (ice): a human splice sites database. *Genomics*, **84**, 762–766.
48. Lee,J., Katari,G. and Sachidanandam,R. (2005) Gobar: a gene ontology based analysis and visualization tool for gene sets. *BMC Bioinformatics*, **6**, 189.
49. Minovitsky,S., Gee,S.L., Schokrpur,S., Dubchak,I. and Conboy,J.G. (2005) The splicing regulatory element, ugcaug, is phylogenetically and spatially conserved in introns that flank tissue-specific alternative exons. *Nucleic Acids Res.*, **33**, 714–724.
50. Sparks,M.E. and Brendel,V. (2005) Incorporation of splice site probability models for non-canonical introns improves gene structure prediction in plants. *Bioinformatics*, **21**, iii20–iii30.
51. Dietrich,R.C., Incorvaia,R. and Padgett,R.A. (1997) Terminal intron dinucleotide sequences do not distinguish between u2- and u12-dependent introns. *Mol. Cell*, **1**, 151–160.
52. Parker,R. and Siliciano,P.G. (1993) Evidence for an essential non-Watson–Crick interaction between the first and last nucleotides of a nuclear pre-mRNA intron. *Nature*, **361**, 660–662.
53. Chanfreau,G., Legrain,P., Dujon,B. and Jacquier,A. (1994) Interaction between the first and last nucleotides of pre-mRNA introns is a determinant of 3′ splice site selection in *S. cerevisiae*. *Nucleic Acids Res.*, **22**, 1981–1987.
54. Deirdre,A., Scadden,J. and Smith,C.W. (1995) Interactions between the terminal bases of mammalian introns are retained in inosine-containing pre-mRNAs. *EMBO J.*, **14**, 3236–3246.
55. Pollard,A.J., Krainer,A.R., Robson,S.C. and Europe-Finner,G.N. (2002) Alternative splicing of the adenylyl cyclase stimulatory G-protein Gα(s) is regulated by SF2/ASF and heterogeneous nuclear ribonucleoprotein A1 (hnRNPA1) and involves the use of an unusual TG 3′-splice site. *J. Biol. Chem.*, **277**, 15241–15251.
56. Schneider,C., Will,C.L., Brosius,J., Frilander,M.J. and Luhrmann,R. (2004) Identification of an evolutionarily divergent U11 small nuclear

ribonucleoprotein particle in *Drosophila*. *Proc. Natl Acad. Sci. USA*, **101**, 9584–9589.
57. Otake,L.R., Scamborova,P., Hashimoto,C. and Steitz,J.A. (2002) The divergent U12-type spliceosome is required for pre-mRNA splicing and is essential for development in *Drosophila*. *Mol. Cell*, **9**, 439–446.
58. Barbosa-Morais,N.L., Carmo-Fonseca,M. and Aparicio,S. (2006) Systematic genome-wide annotation of spliceosomal proteins reveals differential gene family expansion. *Genome Res.*, **16**, 66–77.
59. Wu,H.J., Gaubier-Comella,P., Delseny,M., Grellet,F., Montagu,M.V. and Rouze,R. (1996) Non-canonical introns are at least 10(9) years old. *Nature Genet.*, **14**, 383–384.
60. Wu,Q. and Krainer,A.R. (1999) AT-AC pre-mRNA splicing mechanisms and conservation of minor introns in voltage-gated ion channel genes. *Mol. Cell. Biol.*, **19**, 3225–3236.
61. Lander,E.S., Linton,L.M., Birren,B., Nusbaum,C., Zody,M.C., Baldwin,J., Devon,K., Doyle,M., FitzHugh,W., Funke,R. *et al.* (2001) The international human genome sequencing consortium. Initial sequencing and analysis of the human genome. *Nature*, **409**, 860–921.
62. Montzka,K.A. and Steitz,J.A. (1988) Additional low-abundance human small nuclear ribonucleoproteins: U11, U12, etc. *Proc. Natl Acad. Sci. USA*, **85**, 8885–8889.
63. Tarn,W.Y. and Steitz,J.A. (1996a) Highly diverged U4 and U6 small nuclear RNAs required for splicing rare AT-AC introns. *Science*, **273**, 1824–1832.
64. Patel,A.A., McCarthy,M. and Steitz,J.A. (2002) The splicing of U12-type introns can be a rate-limiting step in gene expression. *EMBO J.*, **21**, 3804–3815.
65. Hertz,G.Z. and Stormo,G.D. (1999) Identifying DNA and protein patterns with statistically significant alignments of multiple sequences. *Bioinformatics*, **15**, 563–577.
66. Schneider,T.D. and Stephens,R.M. (1990) Sequence logos: a new way to display consensus sequences. *Nucleic Acids Res.*, **18**, 6097–6100.
67. Stephens,R.M. and Schneider,T.D. (1992) Features of spliceosome evolution and function inferred from an analysis of the information at human splice sites. *J. Mol. Biol.*, **228**, 1124–1136.
68. Lim,L.P. and Burge,C.B. (2001) A computational analysis of sequence features involved in recognition of short introns. *Proc. Natl Acad. Sci. USA*, **98**, 11193–11198.
69. Carmel,I., Tal,S., Vig,I. and Ast,G. (2004) Comparative analysis detects dependencies among the 5′ splice-site positions. *RNA*, **10**, 828–840.
70. Ast,G. (2004) How did alternative splicing evolve? *Nature Rev. Genet.*, **5**, 773–782.
71. Latijnhouwers,M., Pairoba,C., Brendel,V., Walbot,V. and Urisote,C.-J. (1999) Test of the combinatorial model of intron recognition in a native maize gene. *Plant Mol. Biol.*, **41**, 637–644.
72. Kent,W.J. and Zahler,A.M. (2000) Conservation, regulation, synteny, and introns in a large-scale *C. briggsae-C. elegans* genomic alignment. *Genome Res.*, **10**, 1115–1125.
73. Hollins,C.D., Zorio,D.A., MacMorris,M. and Blumenthal,T. (2005) U2AF binding selects for the high conservation of the *C. elegans* 3′ splice site. *RNA*, **11**, 248–253.