Human Molecular Genetics, 2005, Vol. 14, No. 11 ***1559–1568***
doi:10.1093/hmg/ddi164
Advance Access published on April 27, 2005

Human–mouse comparative analysis reveals that branch-site plasticity contributes to splicing regulation

Guy Kol, Galit Lev-Maor and Gil Ast*

Department of Human Genetics and Molecular Medicine, Sackler School of Medicine, Tel Aviv University, Ramat Aviv 69978, Tel Aviv, Israel

Received February 3, 2005; Revised March 27, 2005; Accepted April 13, 2005

The formation of base-pairing between the branch-site (BS) sequence and the U2 snRNP is an important step in mRNA splicing. We developed a new algorithm to identify both the BS sequence and the polypyrimidine tract (PPT) and validated its predictions experimentally. To assess BS conservation between human and mouse, we assembled and analyzed 46 812 and 242 constitutively and alternatively spliced orthologs of human–mouse intron pairs, respectively. Combinations of BSs and PPTs can be found in most of the constitutive and alternative introns. The average distance between the BS and the 3′ splice site (3′ss) is 33–34 nt. Acceptor-like AG dinucleotides that resided between the predicted BS and the 3′ss were found to appear mostly within 5 nt, but not more than 19 nt, downstream of the BS. However, although 32% of homologous alternatively spliced BS sequences were fully conserved between human and mouse, only a small fraction (3%) of homologous constitutive counterparts was fully conserved. This indicates that the full sequence of the BS is under weak purifying selection in constitutively spliced introns and further strengthens the view that the BS sequence is just one of several factors determining the ability of the splicing machinery to identify the BS location. Mutations in the putative BS revealed a shift from constitutive to alternative splicing, and it also controls the inclusion/skipping ratio in alternative splicing. This suggests a role for BS sequences in regulated splicing.

INTRODUCTION

RNA splicing is a process that removes introns and joins exons to form an mRNA. The splicing mechanism is facilitated by the spliceosome, comprising more than 150 proteins and five complexes of RNA and proteins called snRNPs, marked U1, U2, U4, U5 and U6. The spliceosome catalyzes a two-step enzymatic reaction in which the intron is removed and the two flanking exons are joined. The identification of the correct intron–exon boundaries by the spliceosome depends upon the exact recognition of several *cis*-elements in the pre-mRNA. First, U1 snRNP binds to the 5′ splice site (5′ss), which includes the GT splice donor site in 99.3% of human introns. Also, the splicing factor U2AF65 binds to the PPT, located upstream of the 3′ splice site (3′ss), U2AF35 binds to the 3′ss-AG and SF1 binds to the branch-site (BS) located upstream of the PPT. Next, U2AF65 and SF1 associate with U2 and, presumably, guide its specific binding to the BS area (1–6).

The basal splicing machinery is highly conserved from yeast to human, but the splice site sequences are not (7). The *Saccharomyces cerevisiae* BS consensus sequence (UACUAAC) is fully conserved, and this sequence may be sufficient to guide U2 to the BS (8). In contrast, the mammalian BS is highly variable (8–10); then how is U2 guided to the correct mammalian BS? The answer is partially related to the fact that additional factors such as the RNA structure (11), the PPT-binding protein (12) and the length of the PPT and its C/T content (13) seem to have an effect on U2/BS binding.

Before the first splicing step, U2AF65, U2AF35 and SF1 dissociate from the PPT, 3′ss-AG and BS, respectively. During the second step of splicing, the spliceosome presumably initiates a scanning process aimed at detecting the exact 3′ss-AG (14). Several models were proposed for the scanning process. One

*To whom correspondence should be addressed. Tel: +972 36406893; Fax: +972 36409900; Email: gilast@post.tau.ac.il

model suggested a linear scan, which assigns the first AG downstream of the BS as the 3′ss (15). An important exception to this rule is the case where the AG was <5 nt from the BS; such AG was almost never selected as a 3′ss, probably because of steric effects (15). Another exception accrues when the first AG downstream from the BS (proximal AG) was <6 nt away from an additional AG (distal AG); the distal AG was chosen, indicating competition between adjacent AGs that are <6 nt apart (14). Another important observation was that a proximal AG located <19 nt from the BS is not used efficiently as a 3′ss (16), suggesting that both the distance from the BS to the proximal AG and the space between the proximal and the distal AGs affect the 3′ss selection (16). Also, competition between two AGs leads to the selection of the distal AG as a weak 3′ss that supports alternative splicing (17).

The correct identification of the BS is of critical importance, for better understanding of the splicing mechanism and spliceosome assembly and for screening for mutations that cause aberrant splicing. To this end, several computational methods were developed. Burge *et al.* (18) compiled the consensus sequence of the mammalian BS and Burge and coworkers (19) demonstrated the branch consensus of short introns. However, using the consensus score to locate BS in introns is not sufficient, as sequences with a high score to the consensus have been found to reside in different regions of introns and exons (20). Scanning the 3′ss end of introns, one expects to encounter several sequences with a high score relative to the consensus.

Only a few mammalian BSs have been proven experimentally. Thus, determining a general rule to distinguish between a functional BS and a false consensus hit is very difficult. To improve the BS identification method, we suggest considering the BS sequence, the PPT sequence and the distance between them. It is based on several observations showing that BS selection is affected either by changes in its distance to the PPT (15) or by changes in the PPT length and its C + T content (13).

To determine whether the PPT length and BS-to-PPT distance can be used to locate the functional BS, we collected BS and PPT information from experimentally proven BSs and used it to detect putative functional BSs in human and mouse homologous introns. We showed that putative BSs can be detected in 91% of 46 812 constitutive homologous introns and in 84% of 242 alternative human–mouse homologous introns. Furthermore, AG dinucleotides downstream of those BS that are not used as 3′ss tend to reside mostly within 5 nt, but no more than 19 nt, downstream to the BS—a phenomenon that has been predicted to occur in functional BS. It is interesting to note that only 3% of the constitutive BS sequences are fully conserved in the human–mouse homologous introns, suggesting that the mammalian BS is under low purifying selection. *Ex vivo* mutational analysis of putative BS sequences demonstrated that BS plasticity has a role in alternative splicing.

RESULTS

The validated BS database

The importance of the BS in mRNA splicing has motivated several attempts to define the set of rules that will allow for computational identification of BSs (19,21). However, many mammalian BSs bear little resemblance to conserved sequences that are found in *S. cerevisiae*, which makes this task far more difficult (15). To detect the sequence that might act as a functional BS during splicing, we assembled a data set of 19 experimentally proven BSs (9,2,23). For each BS, we extracted the intron segment ranging from the BS to the 3′ss and calculated the PPT length, which ranged in length from 14 to 33 nt (Fig. 1) (see also Materials and Methods for PPT definition algorithm). For each BS, we calculated the length of the PPT, and in the case of a BS sequence that appeared more than once, we chose the minimal PPT length. This set of intron segments will be called the PROVEN set.

The human–mouse homologous introns database

Approximately 75–130 million years have passed since the human and mouse common ancestor was speciated into separate lineages (24,25). Most of the genes (99%) are orthologs, and the majority of these genes (86%) share the same intron/exon arrangement. Although homologous-exon sequences are highly conserved, intronic counterparts are poorly conserved (7,25–27). Thus, we assumed that the level of conservation of the homologous intronic sequences, like the BS and PPT, would be indicative of their functional importance.

Therefore, we assembled a data set of 93 624 constitutive (CONS) and 484 alternative (ALT) human and mouse homologous introns (see Materials and Methods), respectively. For human and mouse introns that are longer than 300 nt ($n = 73\,090$), the region from −200 to −300 nt upstream of the 3′ss was extracted. This resulted in 73 090 segments (RAND) that were used to represent intronic areas which are not associated with the 3′ end of introns and do not contain functional BSs and PPTs.

Improved method to detect the BS

To evaluate the probability of a search that combines pairs of PPT and BS to identify the location of the functional BS, we first set to search the PROVEN data set for cases of high-score non-BS sequences. We used the BS consensus (18) to score all possible BSs in the last 50 bases of the PROVEN set (see Materials and Methods for the exact scoring algorithm). On the basis of the finding that a 7 nt BS is sufficient to direct U2 binding (9), we calculated the score based on a seven-base motif (5 nt upstream and 1 nt downstream of the branched adenosine, Fig. 1B). Then, we compared the score given to the experimentally proven BS sequence with that of the best-scoring putative BS sequences that were found. In four out of 19 cases (20%), there was another sequence that had a better match with the consensus than the proven BS (marked with dashed underline in Fig. 1A).

Previous works identified possible BSs in introns by searching in the last 50 bases of introns for sequences with high correlation to the BS consensus (19,28). However, this method may be problematic in cases where more than one sequence has a good match to the consensus. The appearance of a high-scoring random sequence is expected to be quite

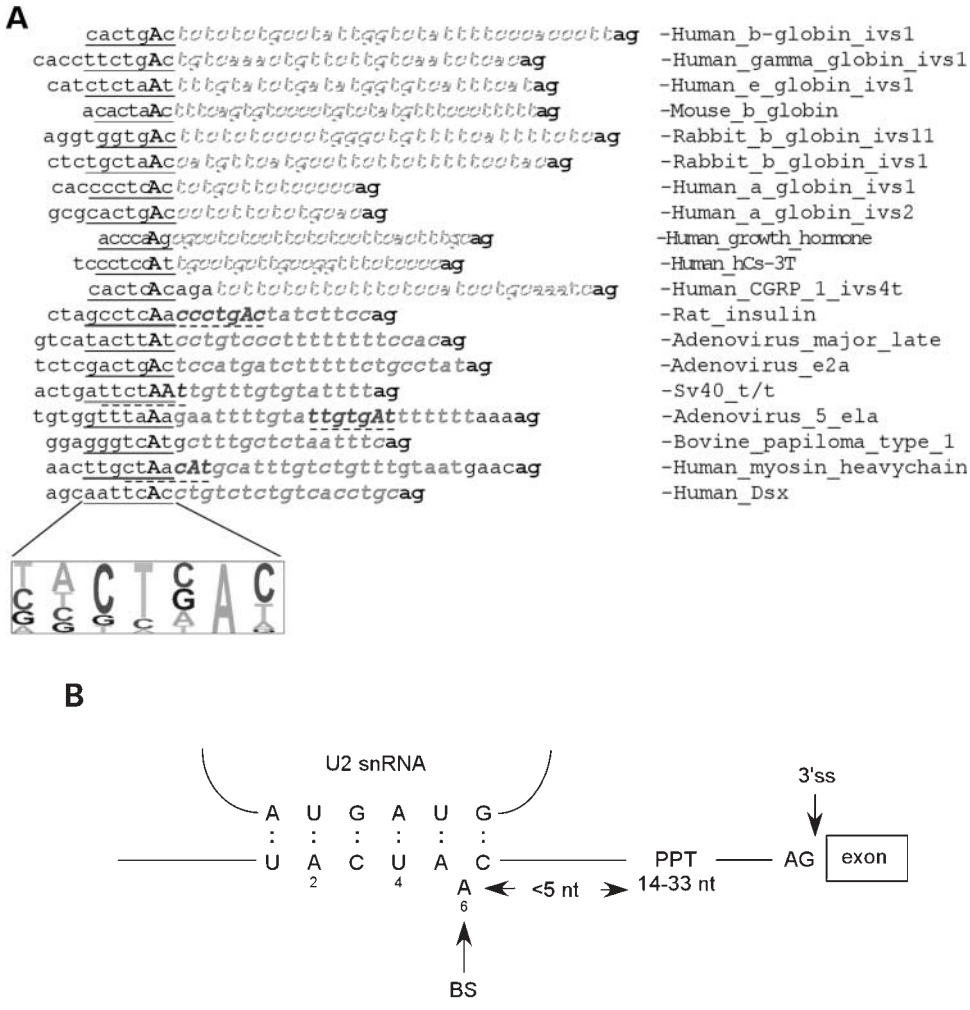

Figure 1. Experimentally validated BS. (**A**) Alignment of the 3′ end of 19 introns in which the BS was experimentally identified. BS area is underlined. The BS adenosine is marked by an uppercase letter. The 3′ss AG is marked by bold letters. The predicted PPT is marked in grey and italic. Dashed underlined sequences show cases in which another part of the sequence had a better match to the BS consensus (19) than the experimentally identified BS. Graphic representation of BS consensus was made using the pictogram program (http:/genes.mit.edu/pictogram.html). (**B**) Diagram of the base pairing between U2 snRNA and the BS sequence. Positions of the BS sequence are indicated. The length of the PPT and the distance between the PPT and the BS in the proven data set is shown in nucleotides (nt).

common, as similar sequences to the BS are known to appear in different intronic and exonic sequences (20).

To estimate how many sequences with a high BS consensus score one could expect to find in an intronic sequence not associated with the 3′ end of introns, we searched the RAND set for hits to the BS consensus. We found that in 60% of the random intronic areas, there is a hit with a score equal to or higher than the minimal score in the PROVEN set (Fig. 2). The same search revealed that 95 and 91% of the CONS and ALT data sets contain a putative BS within the last 100 nt of the intron, respectively. Therefore, these results indicate that when searching for the BS consensus alone we will find, in about 60% of the cases, additional high-scoring BS areas which are not the real BS.

Thus, we decided to find a way to distinguish between functional and random high-scoring sequences. It was shown that, in certain cases, the BS selection depends on the downstream PPT (13,29): mutations that change either the C+T percentage in the PPT or the distance between the BS and the PPT, may cause the splicing system either to select a different BS or to abolish splicing altogether (13,29). On the basis of this mechanism, we hypothesized that when looking for a functional BS one should also consider the composition of the PPT and the distance between the BS and PPT.

Therefore, we searched the introns again, this time allowing only combinations of BS and PPT that are long enough to be functional. First, the search would find BSs with the same sequence of a proven BS that is followed by a PPT at least as long as in the case of the proven BS. As only few BSs were proven experimentally, we decided that in cases in which no proven BS was found we would consider a putative BS (i) if its consensus score was higher than the minimal score of a proven BS and (ii) if the putative BS was followed by a PPT that is longer than the minimal PPT in the PROVEN set (14 nt), residing downstream, but at no more than the maximal observed distance between the BS and the PPT (6 nt).

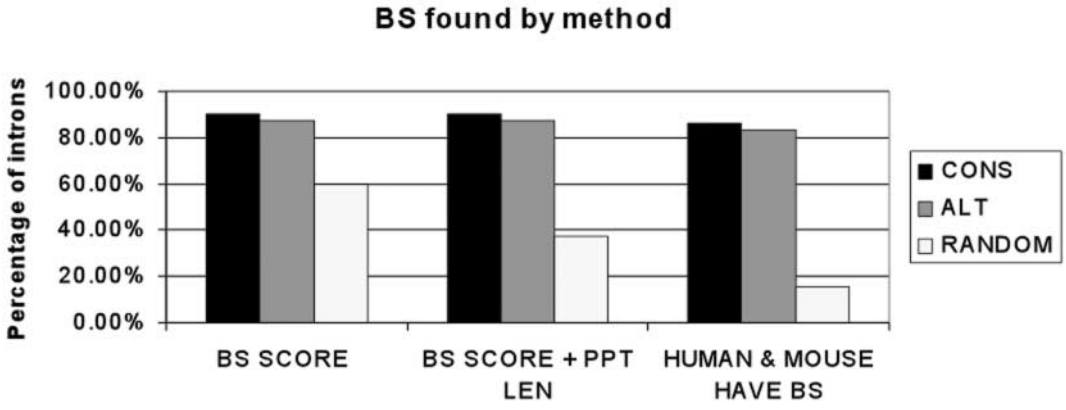

Figure 2. Comparison among different BS recognition methods. Identification of the BS using (i) the consensus sequence only (18) (marked BS SCORE); (ii) the consensus of the validated BS, located at a distance of fewer than 6 bases upstream of a PPT. The PPT should be at least as long as the PPT of the same BS in the VALIDATED BS data set. Alternatively, a BS with a minimal score of BS consensus matrix of Burge *et al.* (18), which is located fewer than six bases upstream of a PPT of 14 nt (marked BS SCORE + PPT LEN) and (iii) cases in which a BS was found independently in both the human and the mouse homologous introns by the BS SCORE + PPT LEN method (marked HUMAN & MOUSE HAVE BS). CONS indicates a data set containing the last 100 bases of 93 624 constitutive introns. ALT indicates a data set containing the last 100 bases of 484 introns preceding alternatively spliced exons. RAND indicates a data set containing 73 090 random intronic sequences (see Materials and Methods for a detailed description of the data set creating methodology).

Using the BS and PPT combinations from the experimentally proven BS set, we first searched the CONS, ALT and RAND sets using the previously mentioned criteria. The results indicate that most of the CONS and ALT BSs were followed by a PPT matching these requirements (Fig. 2; 90 and 88%, respectively). In contrast, only 37% of the RAND introns were found to have a false BS that matched the earlier-mentioned definition. These results show that combinations of BS and PPT that are predicted to support splicing are prevalent in close proximity to the 3′ end of introns rather than in the other intron areas. Also, the average distances between the BS and the 3′ss in the CONS, ALT and RAND sets were 33.09, 34.6 and 33.3, respectively. The similarity of the BS-to-3′ss distance in the RAND data set compared with the CONS and ALT data sets is explained by the relative abundance of poly(T) stretches of lengths of 20–25 nt in intronic sequences originated from retroelements (see also Discussion).

Additional AG downstream to the predicted BS are rare and appear close to the predicted BS

To support the assumption that the predicted BSs found by our analysis are functional, we looked for an additional feature that would characterize the intron area downstream of a functional BS. There are currently two models for the selection of the 3′ss-AG. One suggests that additional AGs, which are not selected as the 3′ss, should be very close to the BS (14); the other suggests that AG dinucleotides that are less than 19 bases downstream of the BS are covered by a protein complex (presumably containing Prp8) and, thus, are not selected as 3′ss (16). The second model suggests that a scanning mechanism starts downstream from this covered area to select the first AG as the 3′ss (16).

Thus, we checked the distribution of AGs (not used as the 3′ss) between the predicted BS and the 3′ss at the CONS, ALT and RAND data sets. For introns that contain one or more additional AGs, we checked the distance from the BS to that AG. Only 14% of the CONS and 17% of the ALT introns contained an additional AG that is not the 3′ss. However, in the RAND set, 66% of the sequences did contain an additional AG between the falsely detected BS and the end of the random sequence. Figure 3 shows the distribution of the distance between the BS and the non-3′ss AG in the case of the CONS and ALT data sets and also the AG that is not at the end of the sequence in the RAND data set. In 70% of the CONS introns, the distance was <5 nt, as opposed to 40% in the RAND data set. The distance distribution shows that the first non-3′ss AG in the CONS tends to appear close to the BS, and the AG in the RAND does not show such a tendency (there were only few cases of non-3′ss AG in the ALT data set—not sufficient for this analysis). We further checked the distribution of the distances between the first and second AGs (not used as 3′ss). In 42% of the CONS introns, the distance was <5 nt, as opposed to 22% in the RAND data set. These results further support the functionality of BS found by our BS search algorithm.

Furthermore, in only 3% of the introns having a non-3′ss-AG downstream of the BS in the CONS data set, this AG is located >19 nt downstream from the BS, with respect to 30% in the RAND set. This result is in agreement with the model suggesting that a protein complex covers the first 19 bases downstream of the BS, and thus the first AG downstream of this area is selected as the 3′ss [unless there is an interplay between two neighbouring AGs (16,17)]. It is also in agreement with the model that suggests that most of the additional AGs appear in close proximity to the BS (<5 nt).

Low level of conservation of BS between human and mouse

Conservation of intron regions between human and mouse is likely to be indicative of functional significance. Thus, we examined the conservation level of those elements between the homologous intron pairs of the CONS and the ALT sets

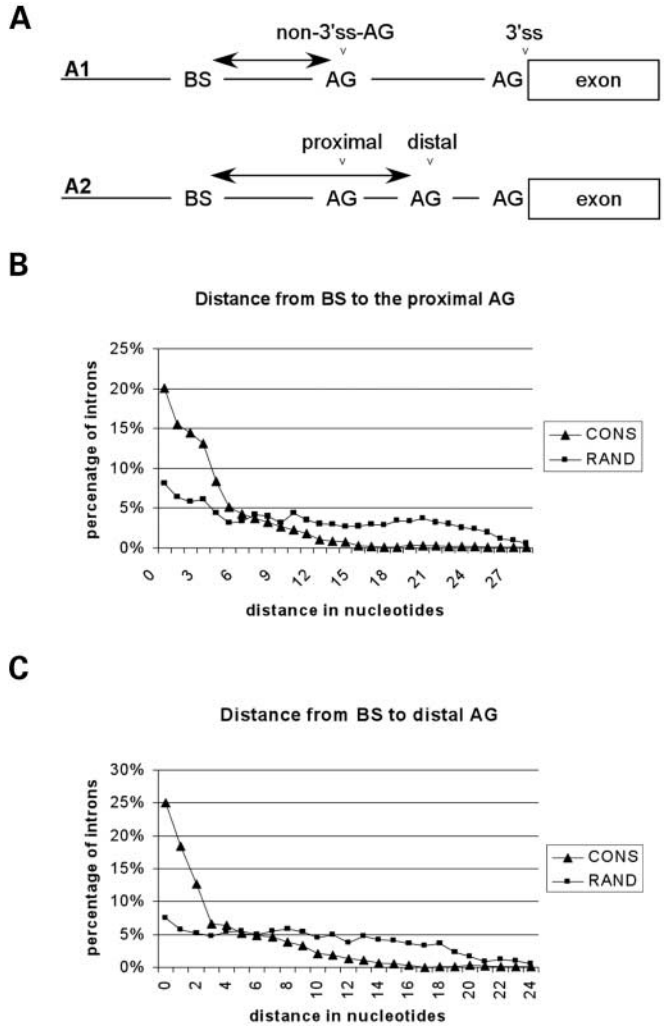

Figure 3. Distance between the BS and the downstream AGs that are not used as the 3′ss. (**A**) Illustration of the distance from BS to first and second AG downstream of the BS, which is not the 3′ss-AG (upper and lower illustrations, respectively). (**B**) Histogram showing the distance between the BS and the downstream non-3′ss AG from 11 326 constitutive introns and 8906 random intronic areas, as illustrated in panel A1. (**C**) The maximal distance to the downstream non-3′ss-AG, as illustrated in panel A2.

and compared that level with the RAND set. We defined a human–mouse intron pair as valid if both the human and the mouse introns have a predicted BS and PPT. In 88 and 68% of the CONS and ALT pairs, respectively, a predictive BS was found with respect to only 15% in the RAND pairs. This indicates that the 3′ ends of human and mouse homologous introns are similar, in the sense that they contain a valid BS and PPT, in most cases, and that this feature is not shared by other parts of the homologous introns (Fig. 2, rightmost panel).

Focusing on the CONS valid pairs, we checked how many of the predicted BS in human–mouse pairs were conserved. We found that only 4% of the human constitutive BS have a possible identical BS sequence in the mouse homologous intron (in the last 100 nt of introns and regardless of the distance of each BS from the 3′ss) when compared with 42% of the alternative introns. When we looked for sequence identity in both BSs, along with identical distance from the 3′ss, the conservation was reduced to 3% in the constitutive introns compared with 32% of the alternative introns. In the random sequences, only 0.44% are fully conserved in sequence and 0.28% in both sequences and distance from the 3′ss. This indicates a very low level of conservation of human–mouse homologous constitutively spliced BS.

We previously demonstrated that intronic sequences located upstream from conserved alternative exons are highly conserved (conservation of ∼100 nt) (30). Therefore, to assess whether the BS conservation is different from the conservation of the area surrounding it, we checked how many of the conserved BS sequences were adjacent to a conserved region of 10 nt directly upstream. We found that 34 and 6% of the ALT and CONS conserved BSs had such a conserved region, respectively. We further checked for conservation of a 7 nt region in the 10 nt region upstream of the conserved BS. We found that 41% of the ALT and 10% of CONS conserved BS had such a conserved region, respectively. This indicates that only a small fraction of constitutive BS sequences is fully conserved, in terms of sequence and distance from the 3′ss. However, the conservation of large intronic sequences flanking alternatively spliced exons suggests that other selective forces affect the BS–PPT region, except for those that maintain the sequence of the BS itself.

PPT length distribution is similar in mouse and human

We next tried to find the reasons for the low level of conservation (3%) of BS located upstream of exons that are constitutively spliced. First, we checked to see whether differences in PPT length cause changes in the distance between the BS and the 3′ss. One possibility might be that neutral mutations in the PPT change its length, without changing the selected BS and the distance between the BS and the 3′ss.

To check this possibility, we first examined whether there is a major difference between human and mouse PPT length in the homologous pairs. Thus, we compared the length of the PPT in the homologous introns of mouse and human. Out of 40 920 human–mouse valid pairs, the human and mouse PPT was longer in 18 869 and 17 386 cases of the human and mouse counterparts, respectively. The mean value of the PPT difference was 0.09 (P-value 0.02 when compared with 0 by t-test). This indicates slightly longer PPTs in human introns with respect to the mouse counterpart.

We next tried to measure the percentage of C+T nucleotide in the PPT of mouse and human. The value of 0.79 in mouse was significantly different from that of the human counterpart of 0.80 (P-value of $1.3 \times E - 04$, Wilcoxon-rank test). To examine whether this difference is caused by a specific position in the PPT, we compared the last 16 positions of the mouse and human introns. No specific position showed a significant difference (data not shown). Thus, we concluded that, although there are differences between human and mouse homologous PPTs, these differences are minor and unlikely to be of biological importance.

A

	Pos 1	Pos 2	Pos 3	Pos 4	Pos 5	Pos 7
Observed	23850	24478	20003	12350	23003	22533
Expected	26624.464	28044.52	24814.71	20680.03	27522.73	27410.45
Obs/Exp	0.8957927	0.872826	0.806094	0.597195	0.835782	0.822059

	Pos 1	Pos 2	Pos 3	Pos 4	Pos 5	Pos 7
Observed	8724	8820	7090	4066	7896	7991
Expected	10901.851	11483.32	10160.82	8467.799	11269.66	11223.69
Obs/Exp	0.800231	0.768071	0.697778	0.480172	0.700642	0.711976

B

		Mouse			
	CONS	A	G	C	T
Human	A	22.38%	7.26%	14.64%	55.71%
	G	13.78%	17.31%	12.90%	56.01%
	C	6.23%	4.27%	29.08%	60.42%
	T	4.12%	2.36%	6.92%	86.61%

Figure 4. Position 4 of the BS is highly conserved. (**A**) Assuming independence among the BS positions, we used the false BS-like areas found in 5885 random intronic human–mouse to represent the background level of changes between non-functional BS. We therefore calculate the expected number of cases in which a specific position will change between those human and mouse random false BS-like sequence. A significant difference in the changing rate of the BS in the CONS and ALT data sets points to evolutionary constraint. To check whether such a difference exists, we then compared the expected percentage of changes to the results obtained from 38 870 cases of human–mouse constitutively spliced homologous pairs. Degree of grey denotes the difference between the expected and observed probabilities. (**B**) The variability degree of specific bases at position 4 in the same data set as in (A). Position 4 was the only position in which the difference between expected and observed changes was statistically significant (53% expected versus 25% observed). The expected probability of specific nucleotide changes was calculated from 5885 cases of random intronic areas of homologous introns in which the BS sequence can be found. For each pair of BS, the changes from the human BS to the mouse BS were recorded. Grey marks cases in which the observed value was at least 2-fold less than the expected.

The second position upstream of the branched adenosine is more conserved than other positions of human–mouse homologous BS

To examine the importance of each position of the BS, we first examined the probabilities of changes per position in a set of BS-like areas that were found via our analysis in the random intronic areas. Next, we compared the changes for each position with the changes observed in the homologous BS sequences. Figure 4A shows a higher conservation of position 4 of the BS (3-fold difference). By assuming independence of the base changes in position 4 of the BS (with respect to mutations in other positions), we checked each possible substitution and compared its observed frequency with the expected one (Fig. 4B). We found a higher level of conservation of the T nucleotide in position 4 of the BS in the homologous BS (88% conservation compared with 58% in RAND). The changes of that T to A, C or G are under purifying selection; T → A, T → G and T → C were 2.1-, 2.5- and 1.76-fold less frequent, respectively, than the changes in the RAND data set. The C in position 4 had a tendency to change to T (66% observed versus 58% expected). The conservation of the T in this position points to the importance of this position in the function of the BS (presumably base pairing with U2) in both human and mouse.

BS mutation promotes a shift from constitutive to alternative splicing

To examine the earlier-mentioned algorithm, we cloned four mini-genes of *Talin, PGT, IMP* and *ADAR2* genes. *Talin, PGT* and *ADAR2* mini-genes are composed from three exons separated by two introns, in which the middle exon is alternatively spliced. The IMP mini-gene is composed of four exons separated by three introns, in which all exons are constitutively spliced. The location of the putative BS sequence in the first intron (for Talin, PGT and ADAR2) was identified using the BS algorithm, and positions 4 and 6 (T and A, respectively) were mutated to G. In IMP, the BS detection algorithm was used to locate the BS in the second intron, and positions 4–7 of the BS were mutated (individually).

Cells were transfected with the mini-genes. After 48 h, cytoplasmic RNA was collected and the splicing pattern was examined using reverse transcription–polymerase chain reaction (RT–PCR) with specific primers to the mini-gene mRNA products. The mutations showed variable effects on splicing from almost complete skipping of the middle exon of Talin and PGT to no effect on the splicing of ADAR2 (compare Fig. 5B, C and E, respectively). In IMP, the mutation of the putative BS-adenosine led exon 12 (third exon in the mini-gene) to shift from constitutive to alternative splicing

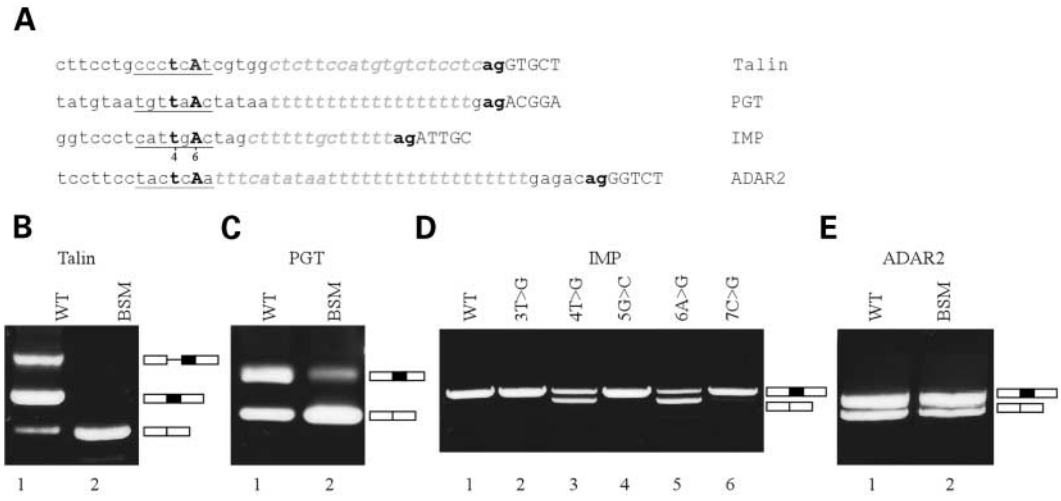

Figure 5. Splicing assays on BS mutants. (**A**) The sequence of the 3′ end of the first intron of Talin, PGT and ADAR2, and the second intron of IMP mini-genes are shown. The positions of the 3′ss-AG, PPT and putative BS are shown in bold, grey and italic and underlined, respectively. The positions that were mutated to G (positions 4 and 6) in Talin, PGT and ADAR2 are in bold, and the mutations in IMP are shown in (D) (**B**–**E**) The indicated plasmid, wild-type or mutant, was introduced into 293T cells by transfection; total cytoplasmic RNA was extracted and splicing products were separated in 2% agarose gel, after reverse transcription polymerase chain reaction (RT–PCR). Lane 1 represents splicing products of wild-type (wt) mini-gene and lane 2 represents splicing products of mutated mini-genes in positions 4 and 6. (D) Lanes 2–6, the indicated mutation in each position is shown in the upper part (refer Fig. 1B for positions of the BS sequence). The mini-gene mRNA isoforms are shown on the right: the black box defines the middle exon and the empty boxes define the flanking exons. The upper mRNA product of Talin contains the first intron (intron retention). All cDNA products were confirmed by DNA sequencing.

(82% exon skipping), compared with 54% exon skipping, when the T in position 4 was mutated to G. Mutations of positions 3, 5 and 7 show no effect on the shift to alternative splicing (Fig. 5D). This further supports the bioinformatics analysis indicating the importance of position 4 of the BS in splicing, but with less weight than that of the BS-adenosine. Although mutation analysis is not a direct proof for BS identification, the finding that mutations of three out of four putative BS showed effects on splicing further supports the validity of this algorithm. The finding that mutations in ADAR2 have no effect on splicing may be due to mis-identification of the correct BS or may be another example of the splicing machinery's ability to select other, non-optimal BSs, in cases where the original BS has been mutated (9).

DISCUSSION

In this report, we have addressed the question of predicting functional BS and PPT residing in human and mouse homologous introns. In addition, we evaluated the conservation level of the BS and PPT between the two organisms. Prediction of BS sequences in mammals is difficult, because BS sequences in mammals are degenerate in nature, and the functional BSs are surrounded by BS-like sequences that are common all along the intron. To overcome this obstacle, we developed a method that uses a combination of PPT length and the BS-to-PPT distance to identify the putative BS. To allow the use of this method to compare the human and mouse BSs, we avoided using conservation information as a parameter in this method.

We applied the method to a data set of 46 812 constitutive and 242 alternative human–mouse homologous intron pairs. This analysis detected putative BS/PPT pairs in 90% of the constitutive and 88% of the alternative introns. The confidence in the functionality of this prediction increases, as there were very few additional AGs between the BS and the 3′ss. Furthermore, in cases where an AG that is not used as the 3′ss does appear between the BS and the 3′ss, the AG appeared close to the BS. In 79% of the cases, this distance is <5 nt, consistent with experimental studies revealing that AGs that are close to the BS are unlikely to act as the 3′ss (14,16). Also, in 99% of the constitutive and alternative BS having additional AGs downstream of the BS, the most distal non-3′ss-AG is no more than 19 nt downstream from the BS. This is in agreement with the model suggesting that a protein complex covers the first 19 nt downstream from the BS (16).

It is interesting that the RAND data set contains 4% of sequences that potentially resemble a BS sequence followed by a PPT. This is especially so when the distance from the false BS sequence to the 3′ end of the RAND sequence is, on average, similar to that of the BS-to-3′ss distance in the CONS and ALT data sets ($\sim$34 nt). This finding is explained by the abundance of retroelements in the human and mouse introns. Common retroelements like the B1 and MIR and human specific retroelements like Alu appear in many copies all along the human and mouse introns. These retroelements contain a poly(A) tail of 20–25 nt, and when inserted in the antisense orientation they create a 20–25 long poly(T) stretch (7). Some of those poly(T) stretches are, by chance, followed by a sequence that resembles a branch point. This abundance of false BS–PPT (some of which might be sites for future exonization of intronic sequences) (7), make the detection of functional BSs even more complex, thus emphasizing the importance of additional parameters (like the AG distribution) to permit distinguishing between real and false BSs.

The conservation of the 3′ intron region among mammals has not been the subject of extensive research as yet. Recent works have demonstrated that 77% of the human–mouse

conserved alternatively spliced exons were flanked on both sides by long, conserved intronic sequences of ~100 nt. However, only 17% of the conserved constitutively spliced exons have a significant alignment of the flanking intronic sequences (30,31). Strikingly, the level of full conservation (both sequence and location) of BS between human and mouse found in our data was very low (3% in constitutive introns, 32% in alternative ones). In 34% of conserved BS in the alternative introns, the BS resides in longer fully conserved areas located upstream of the BS sequence. This was not the case with the conserved BS in the constitutive introns, where only 6% of the conserved BS resided in a longer conserved area extending upstream of the BS sequence. It seems that constitutive BS sequences are under very low purifying selection.

The low level of BS conservation may point to the redundant nature of that sequence in higher eukaryotic cells. Comparative analysis has recently provided insight into the ways in which alternative and constitutive 5′ss vary, giving us hints about the steps involved in the evolution of alternative splicing (7,32,33). We proposed that alternative splicing might have originated as a result of the relaxation of 5′ss recognition in organisms that originally could only support constitutive splicing.

Would the BS plasticity have a similar effect on splicing? One indication that this may be the case is the fact that the BS sequence is highly conserved in yeast *S. cerevisiae*, but it is highly degenerate in higher eukaryotic organisms. This may allow usage of sub-optimal BS sequences in mRNA splicing, thus promoting regulated splicing.

The low conservation level of the BS between human and mouse, the experimental results that show a shift from constitutive to alternative splicing and the effect on the inclusion/skipping ratio in alternative splicing suggest a process in which a highly conserved BS sequence accumulated mutations during evolution. These mutations sub-optimized that site, thus allowing it to be used in alternative splicing as well.

Additional use for the BS identification method may be to detect the location of the putative BS and PPT sequences, when intronic mutations lead to diseases. Diseases caused by mutations that are mapped to putative BS have been demonstrated in genes like *COL51A1* (28), *LDL* receptor (34) and *LCAT* (35). In such cases, the ability to identify the correct BS sequence is of major importance. However, if the specific intron in which the mutation resides includes several putative BSs, identification of the correct BS based solely on the consensus score may lead to incorrect identification of the functional BS, as shown in Figure 1.

In such cases, we suggest using our BS locating method, which searches for a BS consensus located upstream of a PPT. As an additional step, one may pick either only those BSs that are followed by no AG or those for which the additional downstream AG is very close to the BS. For that purpose, we launched a BS search server, using the method described in this manuscript (http://ast.bioinfo.tau.ac.il/). The current BS search is based on the limited set of available experimentally validated cases. Therefore, BS that either are very different from the validated cases or are followed by a very short PPT, will not be identified. However, as most of the human and mouse introns searched in our work did contain a predicted BS and as the distribution of the downstream AG dinucleotides (not used as the 3′ss) matched the definitions obtained from experimentally validated cases, we assume that the method will identify the correct BS in a large fraction of the cases.

MATERIALS AND METHODS

Compiling a database of human–mouse homologous intron pairs

We used a data set of 51 713 exon–intron–exon–intron–exon pairs of human–mouse constitutively spliced exons that we compiled (33). We removed the pairs in which either the human or the mouse intron did not contain GT and AG as the first and last dinucleotides of the introns ($n = 46\,812$). The data set of 242 conserved alternatively spliced intron pairs was obtained from the flanking introns of the conserved human–mouse exons (30).

PPT prediction

The PPT search was first performed on the PROVEN data set. The 50 bases upstream of 3′ss were scanned by a Perl script that located the maximal length subsequence that had the following properties:

(1) The percentage of C+T nucleotides inside the sequence was higher than 0.5. This value was used because PPT of 50% purines was shown to be functional (13).
(2) The 3′ end and 5′ end of this sequence are either C or T.
(3) The most 3′ nucleotides were within 10 nt of the 3′ss. This value was used because the maximal distance from the 3′ss in which a PPT was shown to be functional is 10 nt (36).

If no such sequence was found in the intron, this intron was defined as having no PPT. The minimal length PPT found in the PROVEN data set was of 14 nt. Therefore, when searching for PPT in the CONS, RAND and ALT data sets, we applied the earlier-mentioned set of rules, at the same time considering only PPTs of 14 nt or more to be functional.

BS score calculation

The probability of each nucleotide in a specific position of the BS was adopted from Burge *et al.* (18). We decided to consider only sequences that have an A as the branch nucleotide, because experimental evidence showed that even though C and T can function as the branch nucleotide, A is strongly preferred (37).

Experimental results indicated that a 7 nt BS is enough to support base pairing with U2 (9). Therefore, we decided not to use the whole BS consensus, but, rather, the seven positions spanning from position 3 to position 9.

Assuming independent substitutions of each position, the total score was calculated as the sum of scores of the observed nucleotides over all positions. In mathematical terms, the score calculation was done as follows:

$$\text{Score} = \sum_{j=1\ldots7} \log(\text{prob}(Let, j))$$

where *Let* goes over the set [A,T,C,G] and j marks the position in a sequence of length 7 being scored. Score for the BS was zero for sequences that included letters other than [A,T,C,G].

BS prediction

Assuming that a functional BS must be followed by a functional PPT of a certain length, we used the following method to locate the putative BS.

(1) Scan the 100 nt upstream of the 3′ss and locate all possible PPT sequences, as described in the previous section.
(2) Find possible BSs within the possible distance from a PPT: score all sub-sequences of 7 nt that were within the distance of 6 nt from the 5′ nucleotides of a possible PPT. In the scoring process, we used seven nucleotides, of which five were located upstream and one was located downstream of the BS adenosine. Each subsequence scoring higher than 1.85 was defined as a possible BS. This value was selected because it was the lowest score given to a BS that was validated experimentally.
(3) Discard possible BSs that could not use the PPT that followed them: if the possible BS appeared as a BS in the VALIDATED data set and the possible PPTs were shorter than the PPT found in the VALIDATED data set, this BS was discarded.
(4) Select the highest scoring sequence that survived the earlier steps as the BS.

We have created a public server for BS that will allow BS detection and that also contains the data set: http://ast.bioinfo.tau.ac.il/.

Plasmid constructs

Oligonucleotide primers were designed to amplify four mini-genes: *ADAR2* (adenosine deaminase) contains the exons 7, 8 and 9; *PGT* (putative glycosyltransferase) contains exons 10, 11 and 12; Talin (*TLN*) contains the exons 34, 35 and 36 and *IMP* (IGF-II m-RNA-binding protein) contains exons 10, 11, 12 and 13. Each primer contained an additional sequence encoding a restriction enzyme. The PCR products were restriction designed, and they were inserted between the *Kpn*I–*Bgl*II sites in the pEGFP-C3 plasmid (Clontech).

Site-directed mutagenesis

Oligonucleotide primers containing the desired mutations were used to amplify the mutation-containing replica of the wild-types. The products were treated with *Dpn*I restriction enzyme (12U) (New Englang Biolabs) at 37°C for 1 h. An aliquot of 1–4 μl of the mutant DNA was transformed into *Escherichia coli* DH5α strain. Colonies were picked, followed by mini-prep (QIAgene) and midi-prep (BRL). All plasmids were confirmed by sequencing. The *PGT* mini-gene contains mutations in position 69 and 71 on the intron downstream of the original 5′ss of the Alu antisense sequence, which results in exonization of the intronic Alu in alternative splicing (Ram and Ast, manuscript in preparation).

Transfection, RNA isolation and RT–PCR amplification

The 293T cell line was cultured in Dulbecco's modified Eagle's medium, supplemented with 4.5 g/ml glucose (Renium) and 10% fetal calf serum (Biological industries). Cells were cultured in a 60 mm dish, under standard conditions, at 37°C with 5% CO_2. Cells were grown to 50% confluence and transfection was performed using 12 μl FuGENE6 (Roche) with 4 μg of plasmid DNA. After 48 h, cells were harvested. Total cytoplasmic RNA was extracted using Tri Reagent (Sigma), followed by treatment with 2 U DNase RNase-Free (Ambion). Reverse transcription (RT) was performed on 2 μg total cytoplasmic RNA for 1 h at 42°C, using pEGFP-C3 specific reverse primer and 2 U reverse transctiptase of avian myeloblastosis virus (RT–AMV, Roche). The spliced cDNA products derived from the expressed mini-gene were detected by PCR, using the pEGFP-C3-specific reverse primer and a specific-exon forward primer. Amplification was performed for 30 cycles, consisting of 94°C for 30 s, 61°C for 45 s and 72°C for 1 min, using High Fidelity Taq (Roche). The products were resolved on 2% agarose gel. All PCR products were confirmed by DNA sequencing.

ACKNOWLEDGEMENTS

We thank Ido Carmel, Rotem Sorek, Amir Goren, Noa Sela and Alon Magen for sharing data sets. This work was supported by a grant from the Israel Science Foundation (1449/04 and 717/01) and, in part, by a grant from the German Israeli Project Cooperation Program, FD Hope, and the Chief Scientist of the Israel Health Ministry to G.A.

Conflict of Interest statement. None declared.

REFERENCES

1. Guth, S. and Valcarcel, J. (2000) Kinetic role for mammalian SF1/BBP in spliceosome assembly and function after polypyrimidine tract recognition by U2AF. *J. Biol. Chem.*, **275**, 38059–38066.
2. Liu, Z., Luyten, I., Bottomley, M.J., Messias, A.C., Houngninou-Molango, S., Sprangers, R., Zanier, K., Kramer, A. and Sattler, M. (2001) Structural basis for recognition of the intron branch site RNA by splicing factor 1. *Science*, **294**, 1098–1102.
3. Kielkopf, C.L., Lucke, S. and Green, M.R. (2004) U2AF homology motifs: protein recognition in the RRM world. *Genes Dev.*, **18**, 1513–1526.
4. Brow, D.A. (2002) Allosteric cascade of spliceosome activation. *Annu. Rev. Genet.*, **36**, 333–360.
5. Cartegni, L., Chew, S.L. and A.R. Krainer (2002) Listening to silence and understanding nonsense: exonic mutations that affect splicing. *Nat. Rev. Genet.*, **3**, 285–298.
6. Graveley, B.R. (2001) Alternative splicing: increasing diversity in the proteomic world. *Trends Genet.*, **17**, 100–107.
7. Ast, G. (2004) How did alternative splicing evolve? *Nat. Rev. Genet.*, **5**, 773–782.
8. Yang, C., McPheeters, D.S. and Yu, Y.T. (2004) psi 35 in the branch site recognition region of U2 snRNA is important for pre-mRNA splicing in *Saccharomyces cerevisiae*. *J. Biol. Chem.*, **208**, 6655–6662.

9. Nelson, K.K. and Green, M.R. (1989) Mammalian U2 snRNP has a sequence-specific RNA-binding activity. *Genes Dev.*, **3**, 1562–1571.
10. Berglund, J.A., Chua, K., Abovich, N., Reed, R. and Rosbash, M. (1997) The splicing factor BBP interacts specifically with the pre-mRNA branchpoint sequence UACUAAC. *Cell*, **89**, 781–787.
11. Kent, O.A., Reayi, A., Foong, L., Chilibeck, K.A. and MacMillan, A,M. (2003) Structuring of the 3′ splice site by U2AF65. *J. Biol. Chem.*, **278**, 50572–50577.
12. Lou, H., Helfman, D.M., Gagel, R.F. and Berget, S.M. (1999) Polypyrimidine tract-binding protein positively regulates inclusion of an alternative 3′-terminal exon. *Mol. Cell Biol.*, **19**, 78–85.
13. Norton, P.A. (1994) Polypyrimidine tract sequences direct selection of alternative branch sites and influence protein binding. *Nucleic Acids Res.*, **22**, 3854–3860.
14. 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. *Mol. Cell. Biol.*, **13**, 4939–4952.
15. Smith, C.W., Porro, E.B., Patton, J.G. and Nadal-Ginard, B. (1989) Scanning from an independently specified branch point defines the 3′ splice site of mammalian introns. *Nature*, **342**, 243–247.
16. Chua, K. and Reed, R. (2001) An upstream AG determines whether a downstream AG is selected during catalytic step II of splicing. *Mol. Cell. Biol.*, **21**, 1509–1514.
17. 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*, **300**, 1288–1291.
18. Burge, C.B., Tuschl, T. and Sharp, P.A. (1999) Splicing of precursors to mRNA by the spliceosome. *The RNA World.* Cold Spring Harbor Laboratory Press, Cold Spring Harbor, New York, pp. 525–560.
19. 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.
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. *Methods Enzymol.*, **183**, 252–278.
21. Zhang, M.Q. (1998) Statistical features of human exons and their flanking regions. *Hum. Mol. Genet.*, **7**, 919–932.
22. Zheng, Z.M., Reid, E.S. and Baker, C.C. (2000) Utilization of the bovine papillomavirus type 1 late-stage-specific nucleotide 3605 3′ splice site is modulated by a novel exonic bipartite regulator but not by an intronic purine-rich element. *J. Virol.*, **74**, 10612–10622.
23. Guo, N. and Kawamoto, S. (2000) An intronic downstream enhancer promotes 3′ splice site usage of a neural cell-specific exon. *J. Biol. Chem.*, **275**, 33641–33649.
24. Yang, Z. and Yoder, A.D. (1999) Estimation of the transition/transversion rate bias and species sampling. *J. Mol. Evol.*, **48**, 274–283.
25. Waterston, R.H., Lindblad-Toh, K., Birney, E., Rogers, J., Abril, J.F., Agarwal, P., Agarwala, R., Ainscough, R., Alexandersson, M. and An, P. (2002) Initial sequencing and comparative analysis of the mouse genome. *Nature*, **420**, 520–562.
26. Lander, E.S., Linton, L.M., Birren, B., Nusbaum, C., Zody, M.C., Baldwin, J., Devon, K., Doyle., M. and FitzHugh, W. (2001) Initial sequencing and analysis of the human genome. *Nature*, **409**, 860–921.
27. Modrek, B. and Lee, C. (2002) A genomic view of alternative splicing.*Nat. Genet.*, **30**, 13–19.
28. Burrows, N.P., Nicholls, A.C., Richards, A.J., Luccarini, C., Harrison, J.B., Yates, J.R. and Pope, F.M. (1998) A point mutation in an intronic branch site results in aberrant splicing of COL5A1 and in Ehlers–Danlos syndrome type II in two British families. *Am. J. Hum. Genet.*, **63**, 390–398.
29. Buvoli, M., Mayer, S.A. and Patton, J.G. (1997) Functional crosstalk between exon enhancers, polypyrimidine tracts and branchpoint sequences. *EMBO J.*, **16**, 7174–7183.
30. Sorek, R. and Ast, G. (2003) Intronic sequences flanking alternatively spliced exons are conserved between human and mouse. *Genome Res.*, **13**, 1631–1637.
31. Sugnet, C.W., Kent, W.J., Ares, M., Jr and Haussler, D. (2004) Transcriptome and genome conservation of alternative splicing events in humans and mice. *Pac. Symp. Biocomput.*, 66–77.
32. Sorek, R., Lev-Maor, G., Reznik, M., Dagan, T., Belinky, F., Graur, D. and Ast, G. (2004) Minimal conditions for exonization of intronic sequences; 5′ splice site formation in alu exons. *Mol. Cell*, **14**, 221–231.
33. Carmel, I., Tal, S., Vig, I. and Ast, G. (2004) Comparative analysis detects dependencies among the 5′ splice-site positions. *RNA*, **10**, 828–840.
34. Webb, J.C., Patel, D.D., Shoulders, C.C., Knight, B.L. and Soutar, A.K. (1996) Genetic variation at a splicing branch point in intron 9 of the low density lipoprotein (LDL)-receptor gene: a rare mutation that disrupts mRNA splicing in a patient with familial hypercholesterolaemia and a common polymorphism. *Hum. Mol. Genet.*, **5**, 1325–1331.
35. Li, M. and Pritchard, P.H. (2000) Characterization of the effects of mutations in the putative branchpoint sequence of intron 4 on the splicing within the human lecithin:cholesterol acyltransferase gene. *J. Biol. Chem.*, **275**, 18079–18084.
36. 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.
37. Hartmuth, K. and Barta, A. (1988) Unusual branch point selection in processing of human growth hormone pre-mRNA. *Mol. Cell. Biol.*, **8**, 2011–2020.