4364–4375 *Nucleic Acids Research, 2000, Vol. 28, No. 21*

Analysis of canonical and non-canonical splice sites in mammalian genomes

M. Burset, I. A. Seledtsov and V. V. Solovyev*

Informatic Division, The Sanger Centre, Hinxton, Cambridge, CB10 1SA, UK

Received May 22, 2000; Revised and Accepted August 29, 2000

ABSTRACT

A set of 43 337 splice junction pairs was extracted from mammalian GenBank annotated genes. Expressed sequence tag (EST) sequences support 22 489 of them. Of these, 98.71% contain canonical dinucleotides GT and AG for donor and acceptor sites, respectively; 0.56% hold non-canonical GC-AG splice site pairs; and the remaining 0.73% occurs in a lot of small groups (with a maximum size of 0.05%). Studying these groups we observe that many of them contain splicing dinucleotides shifted from the annotated splice junction by one position. After close examination of such cases we present a new classification consisting of only eight observed types of splice site pairs (out of 256 *a priori* possible combinations). EST alignments allow us to verify the exonic part of the splice sites, but many non-canonical cases may be due to intron sequencing errors. This idea is given substantial support when we compare the sequences of human genes having non-canonical splice sites deposited in GenBank by high throughput genome sequencing projects (HTG). A high proportion (156 out of 171) of the human non-canonical and EST-supported splice site sequences had a clear match in the human HTG. They can be classified after corrections as: 79 GC-AG pairs (of which one was an error that corrected to GC-AG), 61 errors that were corrected to GT-AG canonical pairs, six AT-AC pairs (of which two were errors that corrected to AT-AC), one case was produced from non-existent intron, seven cases were found in HTG that were deposited to GenBank and finally there were only two cases left of supported non-canonical splice sites. If we assume that approximately the same situation is true for the whole set of annotated mammalian non-canonical splice sites, then the 99.24% of splice site pairs should be GT-AG, 0.69% GC-AG, 0.05% AT-AC and finally only 0.02% could consist of other types of non-canonical splice sites. We analyze several characteristics of EST-verified splice sites and build weight matrices for the major groups, which can be incorporated into gene prediction programs. We also present a set of EST-verified canonical splice sites larger by two orders of magnitude than the current one (22 199 entries versus ~600) and finally, a set of 290 EST-supported non-canonical splice sites. Both sets should be significant for future investigations of the splicing mechanism.

INTRODUCTION

Ever since the discovery of split genes it has been observed that practically all introns contain two highly conserved dinucleotides. The donor splice site has GT exactly after the point where the cell cut 5′-end of intron sequences and the acceptor site has AG exactly before the point where the cell cut 3′-end of intron sequences (1,2). With the accumulation of gene sequence data, Mount (3) concluded that this GT-AG rule was always obeyed. However, several cases of splice sites with GC-AG, GG-AG, GT-TG, GT-CG or CT-AG dinucleotides at the splice junctions were observed (4–8). Some of these non-canonical splice sites seemed to be involved in immunoglobulin gene expression and the others in alternative splicing events (6,9).

More recently, a new type of splice pair, the AT-AC, was discovered. It is processed by related, but different, splicing machinery (5,10–18). Introns flanked by the canonical GT-AG pairs are excised from pre-mRNA by the spliceosome including U1, U2, U4/U6 and U5 snRNPs (19). AT-AC introns are excised by a novel type of spliceosome composed of snRNPs U11, U12, U4atac/U6atac and U5 (20–23). Burge *et al.* (24) describe a method to classify splice sites based on the spliceosome machinery involved in their processing and suggest that such splice pairs as AT-AG or AT-AA could represent a transitional state between two splicing systems. However, no systematic analysis of all types of non-canonical splice sites has been carried out.

A draft sequence of the human genome was placed in the public domain recently. However, we lack experimentally acquired knowledge about many genes encoded in this sequence. Therefore the value of the high throughput genomic sequence (HTG) information for the biomedical community will strongly depend on availability of computationally predicted gene candidates. The accuracy of computational

*To whom correspondence should be addressed at present address: EOS Biotechnology, 225A Gateway Boulevard, South San Francisco, CA 94080, USA. Tel: +1 650 246 2331; Fax: +1 650 583 3881; Email: solovyev@eosbiotech.com
Present addresses:
M. Burset, Institut Municipal d'Investigació Mèdica (IMIM), C/Dr Aiguader 80, 08003 Barcelona, Spain
I. A. Seledtsov, Institute of Cytology and Genetics, Novosibirsk, 630090, Russia

gene identification in eukaryotic sequences strongly depends on the splice site models included in the algorithms. Currently, all gene prediction programs generate exon candidates using splice sites with conservative GT and AG dinucleotides. However, ~3.7% of annotated splice sites do not follow this rule and we need to investigate their properties and incorporate this knowledge into gene finding approaches to ensure more accurate genome annotation.

The main goal of this paper is to extract annotated examples of splice sites from the genomic databases and produce an error-free data set of expressed sequence tag (EST)-supported splice pairs. Next, we investigate characteristics of major groups, generate recognition weight matrices or consensus sequences and analyze possible splicing mechanisms based on conserved regions of non-canonical splice sites.

MATERIALS AND METHODS

There are several problems in getting verified information about eukaryotic gene structures from nucleotide sequence databases, such as GenBank or EMBL (25,26). A particular gene could be described in several different entries. For example, information on *Pace4* gene is presented in 17 separate entries that include annotation of uninterrupted sequence regions containing exons and partially sequenced introns. Another problem appears due to annotation errors. It is especially crucial to analyze non-canonical splice sites, because careful checking of 50 such examples revealed 21 cases of clear EMBL annotation errors (27). Therefore, we need to develop a strategy to verify the information presented in the databases.

The problem of selecting all structural information for a given gene from many GenBank entries was solved by the InfoGene database (28), which contains a description of any known gene (positions of exons, introns and alternatively spliced variants) in one entry. We extracted 43 337 pairs of exon–intron boundaries and their sequences from this database covering practically all annotated genes in mammalian genomic sequences.

ESTs are small pieces of mRNA normally of <700 nucleotides and obtained after only one round of sequencing. In principle, the sequences should be obtained only from cytoplasmic extracts, but sometimes they include unprocessed or partially processed transcripts, which could be exported to cytoplasm or produced due to contamination from nuclear RNAs.

To verify extracted splice sites we used alignment of these sequences with known mammalian ESTs (29). In recent years the EST database has grown very significantly and many genes have at least partial information about their expressed regions. Verification of splice sites by ESTs was used earlier in the work of Thanaraj (30), who produced a set of approximately 600 EST-supported canonical splice sites.

We applied EST verification to all pairs of donor and acceptor splice sites including alternative splice variants annotated in the InfoGene database. For each pair we extracted two sequence regions. The first one is 82 nt long sequence around the annotated donor site (40 nt of exon and 42 nt of intron sequence). The conserved dinucleotides of the donor site are exactly centered, with 40 nt on every side. The second 82 nt fragment is the sequence around the acceptor site (42 bp of intron: 2 bp of conserved acceptor dinucleotide and 40 bp of exon sequence).

We will refer to the position typically occupied by GT in donor and by AG in acceptor sites as splice dinucleotides. We also define the exon part of the donor site sequence as exonL (left) and the exon part of acceptor site sequence as exonR (right). We generated a joined sequence for every splice pair combining exonL and the exonR sequences and as a result producing the same sequence as the splicing machinery generated by removing the intron region. We define such 'spliced' sequences of 80 bp (40 bp of exonL and 40 bp of exonR) as spliced constructs (Fig. 1a).

One of our goals was to test whether splice constructs generated from GenBank annotations are actually produced by splicing machinery or whether they are the products of incorrect database annotations. The EST sequence identity with the sequence of splice construct is a good indicator of authenticity of the annotated splice site; especially since the EST data were received independently from the corresponding genomic sequences.

To compare all splice constructs with known EST sequences we obtained an EST data file from NCBI (ftp://ncbi.nlm.nih.gov/blast/db/est.z on August 13, 1999) and ran the blastall program (Version 2.0.9) with the options: 'blastall -p blastn -d est -b 30000 -v 30000 -e 0.01' (31).

As a result, a list of similar ESTs for every spliced construct was generated. Every reported EST can be classified depending on quality and type of observed alignment (Fig. 1b). D-end (donor end coverage): EST partially or totally covers the exonL, but extends in exonR <10 bp (i.e. alignment from 1 to <50 bp of splicing construct). A-end (acceptor end coverage): EST partially or totally covers the exonR, but extends in exonL <10 bp (i.e. alignment from >30 to 80 bp of splicing construct). B-ends (both ends and at least 20 bp match): EST covers (with maximum one substitution) at least the region of 10 bp upstream and downstream from the splice junction (i.e. alignment from 30 to 50 bp of splicing construct) and might be extended in both directions. Error (both ends coverage, error in splice junction): EST has mismatches or gaps in the region of 10 bp upstream or downstream from the splice junction (i.e. alignment from 30 to 50 bp of splicing construct).

When all EST alignments have been classified, every spliced pair can be classified depending on the list of EST alignments with the corresponding splice construct. The examples classified as B-ends, but supported by EST alignments with low conservation (identity is <95%) in 20 bp for every side of splice junction (i.e. from 20 to 60 bp of the splicing construct) were reclassified as unsupported. All other B-ends spliced constructs were considered as supported by ESTs.

In principle, all EST-supported spliced constructs that we obtained after this step should be real, if we assume no errors in considered EST fragments. But this does not imply always that exactly the annotated splice junction is used in splicing. In some cases for non-canonical pairs we can move the annotated position of the splice junction one or more positions upstream or downstream without producing changes in the final splicing construct and encoding the same protein sequence as the annotated one. When such an operation generates a canonical splice site we can consider it as an additional supported splice junction. But we cannot determine which one is real, the generated or annotated. For example, the *Telethonin* gene has only one intron annotated in positions 639–885. This junction

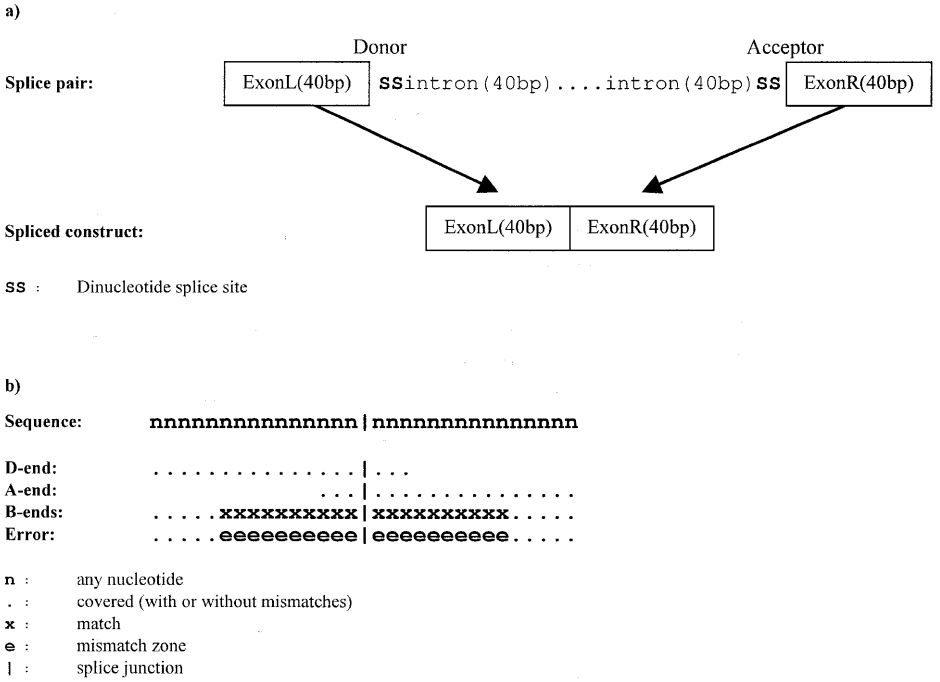


Figure 1. Structure and classification of spliced constructs. (**a**) Structure of spliced constructs. Two sequence regions of a splice pair (marked as Donor and Acceptor) with the corresponding splice site dinucleotides surrounded by 40 bp of gene sequence at each side. Joining exon part of donor (ExonL) and exon part of acceptor (ExonR) we produce a sequence of splice construct to be verified by ESTs. (**b**) EST alignment classification. After obtaining EST and splice construct alignments, every match was classified as D-end (EST covers only the donor part), A-end (EST covers only the acceptor part), B-ends (EST covers a splice junction without mismatches) or Error (EST covers the junction with mismatches).

is completely supported by ESTs, but the annotated splice sites are GG-CA. Analyzing the sequence carefully we found the above-mentioned situation with occurrence of canonical splice pair GT-AG moving the splice junction one position downstream (Fig. 2a). Taking into account that the non-canonical splice junctions occur very rarely, we can suspect that the canonical splice sites are very likely to be the real ones. This observation indicates that ESTs can support an annotated splice junction, but the real junction position can be different. In general, we can think about the possibility of other ESTs crossing the junction, but supporting another splice pair. Again, for such cases we cannot define which one is real. We decided to apply a general approach, generating for every intron pair all possible canonical junctions near the annotated one, in a distance of 10 bp upstream and downstream. We also checked if any of the obtained spliced constructs could be supported by ESTs. We excluded any ambiguous cases from further consideration. In *FUS* gene (Fig. 2b) we have found that we cannot select which one is the correct pair, exactly what was postulated above.

For canonical spliced constructs we applied the same type of analysis. If any annotated canonical junction is supported by ESTs, but we can find another canonical junction also supported by ESTs, then we discard this ambiguous entry because we cannot know which junction is real. The final splice site groups are not only supported by ESTs, but all other possible canonical splice sites near annotated junctions are not supported by ESTs.

RESULTS

We started with 43 337 pairs of donor and acceptor splice sites (splice pairs) from InfoGene database (28). Of these, 1177 are annotated as non-canonical donor sites (2.72%), 993 are annotated as non-canonical acceptor sites (2.29%) and 41 722 (96.27%) contain the canonical splice site pair GT-AG (Table 1). From all splice pairs we generated a table with all possible dinucleotides used in conservative splicing positions and considered the distribution of splice sites among the different groups. The distribution is presented in Table 2 (first number in each cell). Each number in the table shows the absolute number of annotated non-canonical splice pairs (splice pairs with non-canonical conserved dinucleotides in donor or acceptor sites or in both). The last row and column shows the sum for every previous row and column, respectively, to see if some value in a particular line is abnormally high. This table is based on GenBank Release 112. A non-random clustering of 1615 non-canonical splice sites can be clearly seen. Several groups have more than 50 examples and one group contains 245 cases.

Analysis and correction of EST-supported splice junctions

We continue our investigation only for EST-supported sequences, taking into account a significant amount of errors observed in annotated non-canonical splice sites sequences (see Materials and Methods). After analysis of alignments with EST sequences, we found 441 pairs supported by ESTs (27.31%). Interestingly, this percentage is significantly higher for canonical splice pairs. There were 22 374 canonical pairs

Table 1. Statistics of splice pairs

Splice pairs after different filtering stages	Donors	Acceptors	Pairs
Original canonical	42 160 (97.28%)	42 344 (97.71%)	41 722 (96.27%)
Original non-canonical	1177 (2.72%)	993 (2.29%)	1615 (3.73%)
EST-supported canonical	22 437 (98.34%)	22 568 (98.92%)	22 374 (98.07%)
EST-supported non-canonical	378 (1.66%)	247 (1.08%)	441 (1.93%)
EST-supported and corrected canonical	22 252 (98.95%)	22 386 (99.54%)	22 199 (98.71%)
EST-supported and corrected non-canonical	237 (1.05%)	103 (0.46%)	290 (1.29%)
Generalization of analysis of human splice pairs			
GT-AG	22 318	99.24%	
GC-AG	155	0.69%	
AT-AC	12	0.05%	
Other non-canonical	4	0.02%	

supported by ESTs (53.63%). Based on these figures at least half the annotated non-canonical sites should be annotation errors, as was shown in some previous works (5,27). The distribution of EST-supported non-canonical splice sites is presented in Table 2 (second number in each cell). We immediately observe that for a group like GC-AG we have ~50% of EST-supported examples. However, for GT-CA the number is 13 of 66 cases. Probably, groups mostly including true non-canonical pairs (as GC-AG) will have the proportion of EST-supported examples close to the proportion of EST-supported canonical pairs and the groups with a lot of annotation errors will have this value significantly lower (as in GT-TC 26:2; GT-GA 19:1).

The splice pairs, having undetermined position for splice junction (see Materials and Methods; Fig. 2), may be produced by annotation errors or they may produce alternative splicing variants using several EST-supported splice junctions. Due to such ambiguity we have removed 151 non-canonical pairs from our set of analyzed sequences. The distribution of remaining splice pairs among the different groups of conservative dinucleotides is presented in Table 2 (bottom number in each cell). Removing sequences with potential annotation errors in non-canonical EST-supported splice junctions, 290 out of 1615 original pairs remained (17.96%). We call these splice pairs EST-verified.

When we apply the same correction procedure to canonical splice sites, we remove only 175 pairs, where the annotated splice junction positions are ambiguous and could be wrong. Thus, 22 374 out of 41 722 original canonical pairs are supported by ESTs (53.63%) and 22 199 (53.21%) are supported by ESTs after removing those having ambiguous position of splice junctions. Interestingly, this ambiguity is observed for only 0.78% of canonical EST-supported examples, in contrast to the 34% of non-canonical sites. This probably reflects the high error level inside the non-canonical group.

Finally, we are left with 290 non-canonical EST-verified splice pairs per 22 199 EST-verified canonical splice pairs. The proportion of verified canonical splice pairs (98.71%) and non-canonical splice pairs (1.29%) should be closer to the real figures and differ significantly from those found in the original GenBank data (96.27% and 3.73%, respectively).

The 22 199 of verified canonical splice pairs we have deposited to a database (SpliceDB) at http://genomic.sanger.ac.uk . This database contains significantly more verified splice site sequences than the thoroughly investigated set collected earlier (~600 examples) by Thanaraj (30). Our set of 290 EST-verified non-canonical pairs can be used in investigations to verify the reality of these sites, as well as to understand further the splicing machinery. In the SpliceDB description we include the GenBank accession number, intron number, positions of donor and acceptor splice junctions in the InfoGene sequence, sequence around splice sites, type of splice site in the classification presented in this paper as well as the information about EST used to support the splice pair.

Grouping of non-canonical splice pairs

About half (43.45%) of all EST-supported non-canonical splice pairs belong to the GC-AG group (126 members). The next biggest non-canonical group GG-AG is significantly smaller (11). There are many other groups in the same size range, including those processed by the special splicing machinery, the AT-AC group.

The canonical splice sites demonstrate well-defined conserved positions additional to the conserved dinucleotides GT-AG with donor site consensus: AG|GTRAGT and acceptor site consensus: YYTTYYYYYYNCAG|G (32). For the much smaller AT-AC group different conserved positions are found: |ATATCCTTT for donor site and YAC| for acceptor site (20,33,34). These differences reflect some specific interactions with the components of splicing machineries and they can be used to judge if a particular splice site group belongs to the GT-AG or AT-AC splice system.

Examining the sequences of each small group, we find an intriguing feature. In many splice pairs having non-canonical dinucleotides we observe that the canonical splice site conserved dinucleotide has been shifted by one base from the annotated splice junction. For example, in the GG-AG group we can see a typical sequence (underlined): <u>AG</u>|(G)<u>GTAAGT</u>, which is very similar to the canonical consensus, but with a G (in parentheses) inserted between the exonic consensus and the

Table 2. Annotated in GenBank, EST supported and corrected splice site pairs

D\A	AA	AC	AG	AT	CA	CC	CG	CT	GA	GC	GG	GT	TA	TC	TG	TT	Sum
AA	2	1	20	2	6	2	0	0	1	0	2	10	0	0	2	1	49
	0	0	1	1	0	0	0	0	0	0	0	7	0	0	0	0	9
	0	0	1	1	0	0	0	0	0	0	0	1	0	0	0	0	3
AC	1	0	11	2	1	2	0	0	0	1	2	1	0	1	1	1	24
	0	0	0	0	0	1	0	0	0	0	0	1	0	0	0	0	2
	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	1
AG	2	2	34	3	5	4	0	4	1	5	7	3	6	5	7	3	91
	0	2	6	0	1	0	0	2	0	2	4	0	4	1	5	1	28
	0	1	5	0	0	0	0	2	0	1	0	0	0	0	2	0	11
AT	2	20	25	2	1	2	0	0	1	1	2	2	2	0	0	0	60
	1	8	7	2	0	0	0	0	0	1	0	1	0	0	0	0	20
	1	8	7	2	0	0	0	0	0	1	0	1	0	0	0	0	20
CA	1	0	12	0	4	2	1	0	0	0	0	1	0	1	3	2	27
	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	1	2
	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	1	2
CC	0	1	8	1	2	1	0	0	1	1	1	1	1	0	1	1	20
	0	0	2	0	0	0	0	0	0	0	0	0	0	0	0	0	2
	0	0	2	0	0	0	0	0	0	0	0	0	0	0	0	0	2
CG	0	0	17	0	5	0	0	0	0	0	1	1	0	0	0	0	24
	0	0	1	0	1	0	0	0	0	0	0	0	0	0	0	0	2
	0	0	1	0	1	0	0	0	0	0	0	0	0	0	0	0	2
CT	0	13	19	0	2	2	0	2	0	1	0	0	0	1	2	0	42
	0	2	6	0	1	0	0	0	0	0	0	0	0	0	0	0	9
	0	2	6	0	1	0	0	0	0	0	0	0	0	0	0	0	9
GA	1	4	41	0	8	0	0	1	2	3	5	14	0	2	3	0	84
	0	0	8	0	0	0	0	0	0	1	0	7	0	1	1	0	18
	0	0	8	0	0	0	0	0	0	0	0	1	0	1	1	0	11
GC	0	0	245	0	4	1	0	1	2	1	5	0	1	0	2	1	263
	0	0	126	0	0	0	0	0	0	0	1	0	1	0	0	0	128
	0	0	126	0	0	0	0	0	0	0	1	0	1	0	0	0	128
GG	5	7	76	5	68	0	1	3	6	3	5	5	19	2	2	3	210
	0	1	12	0	41	0	0	0	2	1	0	0	6	2	0	0	65
	0	1	11	0	5	0	0	0	2	0	0	0	0	2	0	0	21
GT	26	21	*	17	66	49	20	15	19	22	50	16	23	26	38	30	438
	0	4		2	13	2	4	3	1	1	11	1	7	2	10	2	63
	0	4		2	9	0	4	3	0	1	10	1	7	2	8	2	53
TA	0	3	32	1	8	1	2	2	12	4	38	5	1	1	0	2	112
	0	0	8	0	1	0	1	0	5	2	23	1	0	1	0	0	42
	0	0	6	0	0	0	1	0	0	0	0	0	0	1	0	0	8
TC	0	0	8	0	0	0	0	0	0	0	5	1	0	1	0	0	15
	0	0	1	0	0	0	0	0	0	0	2	1	0	0	0	0	4
	0	0	1	0	0	0	0	0	0	0	1	0	0	0	0	0	2
TG	0	2	50	0	9	2	1	2	8	4	27	1	1	0	1	2	110
	0	1	10	0	1	0	0	0	1	1	17	0	0	0	1	0	32
	0	1	7	0	0	0	0	0	0	0	2	0	0	0	0	0	10
TT	1	0	24	1	1	1	0	1	1	2	10	1	2	0	0	1	46
	0	0	5	1	0	0	0	0	1	0	8	0	0	0	0	0	15
	0	0	5	1	0	0	0	0	0	0	1	0	0	0	0	0	7
Sum	41	74	622	34	190	69	25	31	54	48	160	62	56	40	62	47	1615
	1	18	194	6	59	3	5	5	10	9	66	19	18	7	17	4	441
	1	17	187	6	16	1	5	5	2	3	15	4	8	6	11	3	290

First number in every cell shows the number of GenBank annotated pairs, second number is EST-supported pairs and last number means supported by ESTs and corrected, as explained in the text.
D, donor; A, acceptor.
*41 722, 22 374 and 22 199 are canonical splice pairs.

intronic consensus parts (compared with the canonical consensus: AG|GTRAGT).

The approach taken is to analyze the instances of standard dinucleotides upstream or downstream of splice junction in each verified non-canonical splice site group with more than three representatives (17 non-canonical groups). For example, doing this for 11 verified GG-AG pairs we observed that nine pairs have shifted canonical donor splice site (GT dinucleotides), one major non-canonical site with GC dinucleotide and one GA-AG case (Fig. 3).

Such 'shifted' cases were found in 10 out of 17 non-canonical groups. Notably, all but two cases (discussed below) can be reclassified as belonging to one of seven non-shifted groups or to the canonical pairs [when a standard dinucleotide shifted by one base from the annotated splice junction is taking place (Fig. 4a)]. This grouping makes no apparent biological sense, but it restricts the non-canonical splice groups to eight possible types, producing a rather symmetrical classification (Fig. 4b). In the first exception example (of two unclassified cases) we observe the annotated genomic junction ACC|ctgc…ggag|CTG. In ESTs supporting this junction we found ACG|CTG, i.e. a substitution in the last donor exonic nucleotide. Checking donor –1 position allowed us to recover a GC-AG pair. The second exception presents EST supported splice junction that is annotated as TGG|gggt…ttca|GCT. However, the same EST will equally support TGGG|ggt…ttcag|CT, also, which contains shifted canonical pair: TGGG(G)|gt…ttcag|CT.

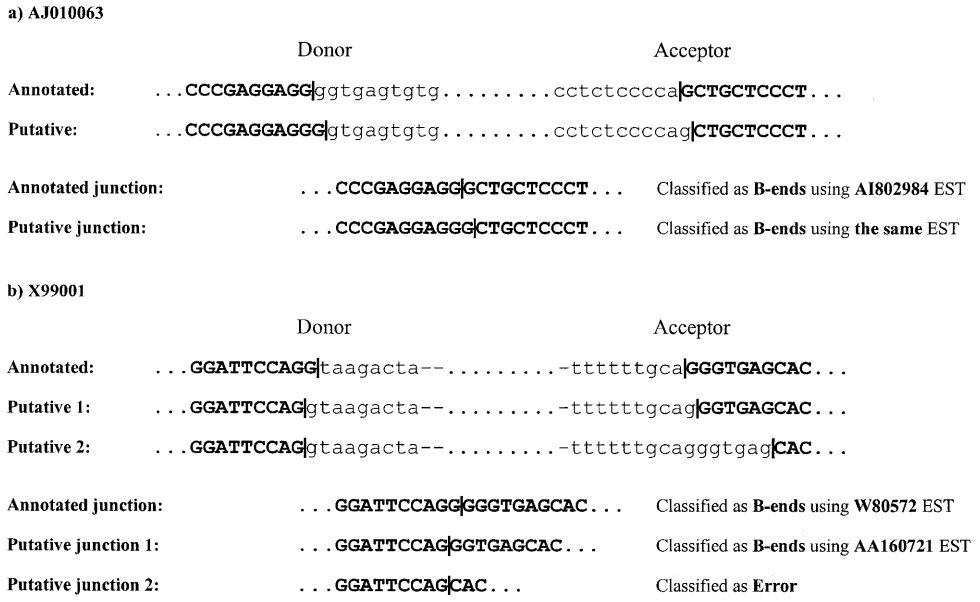


Figure 2. Examples of possible ambiguities in supported by EST splice pairs. (**a**) *Homo sapiens Telethonin* gene, intron 1 (AJ010063). An example of annotated non-canonical junction supported by EST. The same EST can also support a canonical splice junction. The annotated non-canonical junction and the putative canonical one produce the same spliced sequence. (**b**) *Homo sapiens FUS* gene, intron 14 (X99001). An example of annotated non-canonical junction supported by EST. Another EST supports a closely located canonical splice junction. In this case the EST-supported putative spliced sequence differs by 2 nucleotides (gg) from the annotated one.

	Donor	Acceptor	Donor +1 shift
U07083	**AAGGG**|ggtaagg	ctttaag|**GGTGT**	GG ⇒ GT
L43831	**CAAAG**|ggtactg	tctgcag|**CTTTG**	GG ⇒ GT
U37431	**AAACA**|ggtcagt	gccccag|**GGGAA**	GG ⇒ GT
U02978	**AGGCC**|ggtgagt	gggccag|**GGGTC**	GG ⇒ GT
AJ000060	**AGTAT**|ggtaagg	tttccag|**GGAGA**	GG ⇒ GT
U12599	**GCTGG**|ggtaagt	tccccag|**TCATA**	GG ⇒ GT
U01247	**TCACA**|ggtatgc	attctag|**GAGAA**	GG ⇒ GT
U28721	**CGCAG**|ggcaagg	ctaacag|**GTCTA**	GG ⇒ **GC**
M20214	**AACAG**|ggaaggc	acgctag|**GGAAA**	GG ⇒ **GA**
M62601	**TGCAG**|ggtatac	cctttag|**ACAAT**	GG ⇒ GT
U66878	**TAGTG**|ggtgagt	ccttcag|**GAGTG**	GG ⇒ GT

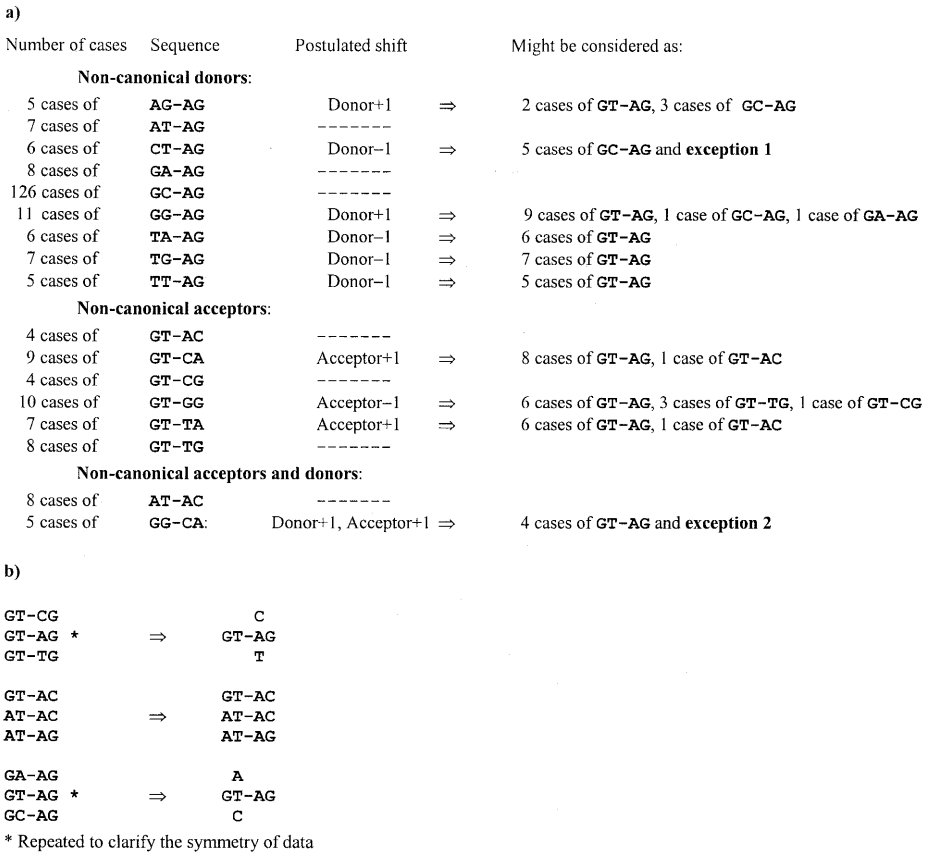
Figure 3. Shifted splice sites. Examples of GG-AG verified splice pairs (11 cases). In donor sites (exactly after the cut point) a GG pair is always found. To decide to which type of splicing pair we should assign these non-canonical examples we checked all closely located standard dinucleotides. They are found shifted by 1 nucleotide downstream. We reclassify the presented splice pairs as nine canonical GT-AG, one GC-AG and one GA-AG site.

Applying HTG sequences for verification of non-canonical splice pairs

The most likely explanation of shifted canonical dinucleotides near the non-canonical splice sites is sequence annotation errors, as someone inserted/deleted one additional nucleotide, that is actually absent/present in real genomic sequence. EST alignments allow us to verify the position of splice junction and the sequence in exons, but we have no additional information for the intron sequence, in which splicing conserved dinucleotides are localized. To check sequences in intronic regions we need an independent source of genomic information. To address this question we decided to compare our EST-verified non-canonical pairs with HTG sequences. From the GenBank sequence server (ftp://ncbi.nlm.nih.gov/genbank/) we have retrieved the file with human sequences and studied all HTG-supported human non-canonical splice pairs. For every human EST-verified splice pair (171 entries) we prepared two sequences, a 40 bp sequence upstream and a 40 bp sequence downstream of donor site dinucleotides and the same was done for acceptor sites. The HTG sequences were converted to a BLAST database. Aligning selected non-canonical donor and acceptor sequences that allowed us to recover all human splicing pairs present in GenBank and high throughput human sequences simultaneously; 156 cases were found (interestingly 91% of human sequences had matches in HTG). A summary of this analysis is presented in Figure 5.

The results show that practically all human EST-supported GC-AG cases were supported by the HTG matches (78 of 79 cases) and additionally we recovered one GC-AG case from an error in the annotated splice site. We found 53 other errors that damage canonical splice pairs. One type of error includes cases in which intronic GenBank sequences are completely absent in the corresponding HTG sequences. Other cases had intronic GenBank sequences with small gaps or substitutions in exonic and intronic parts. Two non-canonical sites (from the same entry) were annotated incorrectly in the forward DNA chain. Analyzing the bibliography information cited in GenBank (35,36) we found that they actually occur in the reverse chain and both sites are canonical. Additionally, we found six cases of annotated pseudogenes that will be studied in detail below. We identified six AT-AC pairs (four pairs were correctly annotated in the original non-canonical set and two were recovered from errors; Fig. 5). Beside that, one case was annotated as intron, but in HTG the exonic parts were continuous. Seven cases of HTG were identical to GenBank sequences and, for this reason, excluded from the analysis.

Finally, we obtained only two non-canonical pairs that were supported by EST and HTG sequences. The first one (U01337)



a)

Number of cases	Sequence	Postulated shift		Might be considered as:
Non-canonical donors:				
5 cases of	AG-AG	Donor+1	⇒	2 cases of GT-AG, 3 cases of GC-AG
7 cases of	AT-AG	-------		
6 cases of	CT-AG	Donor–1	⇒	5 cases of GC-AG and **exception 1**
8 cases of	GA-AG	-------		
126 cases of	GC-AG	-------		
11 cases of	GG-AG	Donor+1	⇒	9 cases of GT-AG, 1 case of GC-AG, 1 case of GA-AG
6 cases of	TA-AG	Donor–1	⇒	6 cases of GT-AG
7 cases of	TG-AG	Donor–1	⇒	7 cases of GT-AG
5 cases of	TT-AG	Donor–1	⇒	5 cases of GT-AG
Non-canonical acceptors:				
4 cases of	GT-AC	-------		
9 cases of	GT-CA	Acceptor+1	⇒	8 cases of GT-AG, 1 case of GT-AC
4 cases of	GT-CG	-------		
10 cases of	GT-GG	Acceptor–1	⇒	6 cases of GT-AG, 3 cases of GT-TG, 1 case of GT-CG
7 cases of	GT-TA	Acceptor+1	⇒	6 cases of GT-AG, 1 case of GT-AC
8 cases of	GT-TG	-------		
Non-canonical acceptors and donors:				
8 cases of	AT-AC	-------		
5 cases of	GG-CA:	Donor+1, Acceptor+1	⇒	4 cases of GT-AG and **exception 2**

b)

```
GT-CG                    C
GT-AG *        ⇒      GT-AG
GT-TG                    T

GT-AC                  GT-AC
AT-AC          ⇒       AT-AC
AT-AG                  AT-AG

GA-AG                    A
GT-AG *        ⇒      GT-AG
GC-AG                    C
```

* Repeated to clarify the symmetry of data

Figure 4. Analysis of EST-supported non-canonical splice site groups. (**a**) Classification. Analyzing all EST-verified non-canonical splice pairs and taking into account cases with shifted canonical consensus this classification has been produced. Practically all splice pairs have only one non-canonical splice dinucleotide. (**b**) Table of possible splice pairs. After generalization we have obtained only seven non-canonical splice pair groups and a total of eight groups if we include the canonical splice pairs. The first (top) part of the right figures shows canonical donor site combined with all observed variations of acceptor site (GT-AG, GT-CG and GT-TG). The second (middle) part shows AT-AC group and hybrid pairs (GT-AC, AT-AC and AT-AG). The third (bottom) part shows canonical acceptor site combined with all observed variations of donor site (GA-AG, GC-AG and GT-AG).

has a GT-GG pair and is annotated with a description of cDNA for this gene (37). The second pair (AF109620) has a TT-AG pair and comes from an entry with partially sequenced introns. There is no bibliographic reference for this example (direct submission).

We have commented above that six cases resemble pseudogene sequences (Fig. 5b). For each of these pairs several HTGs (with different locations in genome) were identified, when we compared the GenBank sequence with the HTG set. One HTG sequence had a non-canonical pair (the same as the GenBank sequence) in the corresponding position, but another had a canonical pair. To define which genomic sequence corresponds to the functional gene we investigated differences in exonic parts of these sequences close to splice junction. We assumed that only functional (i.e. expressed) genomic sequence should coincide perfectly with the ESTs in these positions. Comparing two HTG sequences highly homologous to the annotated non-canonical D87002 splice pair, we found that all ESTs (more than five) support the substitution from A to G (upstream the donor site), which occurs only in HTG sequence with canonical splice pair. In the X14615 splice pair we found a similar situation, a substitution from T to C downstream of the acceptor site. U41163 example is also very similar, it has two pseudogene sequences and the functional has the shifted donor splice junction by one position (relative to the GenBank report). In this case the canonical site is EST-supported having the substitution G to C downstream of the acceptor site. The X72812 case is practically the same as X02725, differing in only one substitution located 34 bp upstream of the donor site, so in the region of splice junction these examples are the same, and both of them are annotated as immunoglobulin kappa light chain variable region. This is an interesting case of pseudogene, because we can see at least three different copies in HTG. Based on the differences in exonic parts with respect to ESTs, we can detect which copy is functional. Fortunately, we have more than five ESTs supporting two changes compared with the annotated GenBank sequence. The first one is a substitution (A to G) 9 bp downstream of the acceptor junction. The second is two substitutions downstream of the acceptor site (Fig. 5b). It should be noted that one of the pseudogenes maintains a conserved canonical splice site pair. Finally, the strangest example of pseudogenes appearing in this study is M71243. We recovered four different fragments from HTG having more than five ESTs supporting the junction. We observed that in all four HTG cases there are substitutions in the same positions in ESTs (upstream donor in 2, 7 and 10; and downstream acceptor in

a) General results.

290 EST-verified non-canonical splice sites
171 human EST-verified non-canonical splice sites
156 human EST-verified non-canonical splice sites with matches in HTG, of which:
- 55 GT-AG cases (all corrected)
- 79 GC-AG cases (78 annotated + 1 corrected)
- 6 AT-AC cases (4 annotated + 2 corrected **(b)**)
- 7 HTG entered in GenBank
- 1 Non-existent introns **(b)**
- 6 Annotated pseudogenes **(b)**
- 2 non-canonical **(c)**

b) Interesting errors.	**Donor**	**Acceptor**	**ESTs**	**NumESTs**
Corrected AT-AC cases				
AC006942 (Alpha adaptin A)				
GenBank:	**CCAGACTCAT**|atcctctcag	gcgggcacag|**CTGGCTGTGC**		
High througput:	**CCAGACTCAT**|atcctctcag	gcgggcacag|**CTGGCTGTGC**		
ESTs	**CCAGACTC**|	|**AGCTGGCTGTGC**	T78562	3
U53331 (spermidine apt.)				
GenBank:	**TGGTAGAGAT**|atcctttgtt	ttgcggacat|**TGACCAAATG**		
High througput:	**TGGTAGAGAT**|atcctttgtt	ttgcggacat|**TGACCAAATG**		
ESTs	**TGGTAGAG**|	|**ATTGACCAAATG**	AA136221	5+
Annotated pseudogenes				
D87002 (gamma-glut. tranp.)				
GenBank:	**A..CACTGCACTG**|atatgtgtca	catgccccag|**GCCATCATCT**		
High througput:	**A..CACTGCACTG**|atatgtgtca	catgccccag|**GCCATCATCT** (pseudo)		
High througput:	**G..CACTGCACTG**|gtatgtgtca	catgccccag|**GCCATCATCT**		
ESTs	**G..CACTGCACTG**|	|**GCCATCATCT**	AI807626	5+
X14615 (Ig variable chain lambda)				
GenBank:	**CACTGCACAG**|gtgcccagac	ctgcttccac|**GGTCCTGGGC..T**		
High througput:	**CACTGCACAG**|gtgcccagac	ctgcttccac|**GGTCCTGGGC..T** (pseudo)		
High througput:	**CACTGCACAG**|gtgcccagac	ctgcttccag|**GGTCCTGGGC..C**		
ESTs	**CACTGCACAG**|	|**GGTCCTGGGC..C**	AA482644	5+
U41163 (Creatine transporter)				
GenBank:	**CTGCTACAAG**|taagcactgc	acccctccag|**GACGCCATCA..G**		
High througput:	**CTGCTACAAG**|taagcactgc	acccctccag|**GACGCCATCA..G** (pseudo)		
High througput:	**CTGCTACAAG**|taagcactgc	acccctccgg|**GACGCCATCA..G** (pseudo)		
High througput:	**CTGCTACAA**|gtaagcaccgc	acccctcag|**GGACGCCATCA..C**		
ESTs	**CTGCTACAA**|	|**GGACGCCATCA..C**	H82584	5
X72812 (Ig kappa light chain variable region L10) and X02725 (Ig kappa light chain variable region V(h))				
GenBank:	**TGGCTCCCAG**|gtgaggggaa	ccaatcttgg|**ATACCACCAG..AA**		
High througput:	**TGGCTCCCAG**|gtgaggggaa	ccaatcttgg|**ATACCACCGG..AA** (pseudo)		
High througput:	**TGGCTCCCAG**|gtgaggggaa	ccaatctcag|**ATACCACCGG..AA** (pseudo)		
High througput:	**TGGCTCCCAG**|gtgaggggaa	ccaatttcag|**ATACCACCGG..GT**		
ESTs	**TGGCTCCCAG**|	|**ATACCACCGG..GT**	AA402152	5+
M71243 (glycophorin Sta type A)				
GenBank:	**AAGAAAAACG**|ttatgttctt	tgctttatag|**GAGAAAGGGT**		
High througput:	**AAGAAAAACG**|ttatgttctt	tgctttatag|**GAGAAACGGG** (pseudo)		
High througput:	**GAGGAAAATG**|atatgttctt	cgctttatag|**GAGAAAGGGG** (pseudo)		
High througput:	**GAGGAAACCG**|gtatgttctt	cgctttatag|**GAGAAACGGG** (pseudo?)		
High througput:	**GAGGAAACCG**|gtatgt-ctt	cgctttatag|**GAGAAACGGG** (pseudo?)		
ESTs	**GAGGAAACCG**|	|**GAGAAAGGGT**	T95140	5+
Non existent introns				
M13300 (RCP)				
GenBank:	**CCCACTCGCT**|atcctttgtt	tctcccttag|**ATCATCATGC**		
High througput:	**CCCACTCGCT**|atcctttgtt	cccactcgct|**ATCATCATGC**		
ESTs	**CCCACTCGNT**|	|**ATCATCATGC**	T27896	2
ESTs	**CCCACTCAGC**|	|**ATCATCATGC**	W28410	3

c) Supported human HTG sites.	**Donor**	**Acceptor**	**ESTs**	**NumESTs**
U01337 (A-RAF-1)				
GenBank:	**AAGTCTAACA**|gtatctatct	acc-tggcgg|**ACATCTTCCT**		
High througput:	**AAGTCTAACA**|gtatctatct	accctggcgg|**ACATCTTCCT**		
ESTs	**AAGTCTAACA**|	|**ACATCTTCCT**	F13363	3
AF109620 (ETV1)				
GenBank:	**AGAACAGAAG**|ttaagttgtc	tcatccacag|**GCTGTATGTT**		
High througput:	**AGAACAGAAG**|ttaagttgtc	tcatccacag|**GCTGTATGTT**		
ESTs	**AGAACAGAAG**|	|**GCTGTATGTT**	T19167	1

Figure 5. Comparison of human GenBank sequences and available HTGs.

7 and 10). They are not present completely in any HTG fragment, so in principle we should consider all four as possible pseudogenes. Probably the gene with the functional splice pair has not been cloned yet. Two out of four of these sequences contain canonical splice pairs, so we can consider (as for X72812) that observed pseudogenes are supporting a canonical pair.

Table 3. Characteristics of major splice pair groups

GT-AG group[a]															
Donor frequency matrix															
A	34.0	60.4	9.2	0.0	0.0	52.6	71.3	7.1	16.0						
C	36.3	12.9	3.3	0.0	0.0	2.8	7.6	5.5	16.5						
G	18.3	12.5	80.3	100	0.0	41.9	11.8	81.4	20.9						
U	11.4	14.2	7.3	0.0	100	2.5	9.3	5.9	46.2						
Acceptor frequency matrix															
A	9.0	8.4	7.5	6.8	7.6	8.0	9.7	9.2	7.6	7.8	23.7	4.2	100	0.0	23.9
C	31.0	31.0	30.7	29.3	32.6	33.0	37.3	38.5	41.0	35.2	30.9	70.8	0.0	0.0	13.8
G	12.5	11.5	10.6	10.4	11.0	11.3	11.3	8.5	6.6	6.4	21.2	0.3	0.0	100	52.0
U	42.3	44.0	47.0	49.4	47.1	46.3	40.8	42.9	44.5	50.4	24.0	24.6	0.0	0.0	10.4
GC-AG group[b]															
Donor frequency matrix															
A	40.5	88.9	1.6	0.0	0.0	87.3	84.1	1.6	7.9						
C	42.1	0.8	0.8	0.0	100	0.0	3.2	0.8	11.9						
G	15.9	1.6	97.6	100	0.0	12.7	6.3	96.8	9.5						
U	1.6	8.7	0.0	0.0	0.0	0.0	6.3	0.8	70.6						
Acceptor frequency matrix															
A	11.1	12.7	3.2	4.8	12.7	8.7	16.7	16.7	12.7	9.5	26.2	6.3	100	0.0	21.4
C	36.5	30.9	19.1	23.0	34.9	39.7	34.9	40.5	40.5	36.5	33.3	68.2	0.0	0.0	7.9
G	9.5	10.3	15.1	12.7	8.7	9.5	16.7	4.8	2.4	6.3	13.5	0.0	0.0	100	62.7
U	38.9	41.3	58.7	55.6	42.1	40.5	30.9	37.3	44.4	47.6	27.0	25.4	0.0	0.0	7.9

Frequencies of bases in percents for every significant position around donor and acceptor sites in GT-AG (canonical) and GC-AG pairs.
[a]Number of EST-supported cases: 22 199. Frequency: 22 199/22 489 (98.71%).
[b]Number of EST-supported cases: 126. Frequency: 126/22 489 (0.56%).

Estimation of frequency of non-canonical splice sites

If we assume that approximately the same situation is true for the whole set of annotated mammalian non-canonical splice sites, then we could compute expected frequencies of every possible splice pair. Before further calculations are carried out, we should subtract seven HTG sequences included in GenBank and the case of no-intron sequence from 156 supported by HTG human sequences. Thus, the number of verified examples is 148, representing 51% of the total non-canonical EST-supported sequences. We should take this into account when extrapolating the results to the whole set of mammalian non-canonical splice pairs.

Following this procedure we obtained 79 out of 148 human GC-AG pairs supported by HTG sequences. Extrapolating this to the whole set of non-canonical splice sites (290 cases) we predict 155 GC-AG pairs. That number is similar to the real number of observed cases (126). The proportion of GC-AG in respect to the total splice site pairs (155/22 489 × 100) is 0.69%. Using the same procedure with 61 recovered standard pairs we estimate their number among the annotated non-canonical pairs as 119 cases. The percent of standard GT-AG pairs [(22 199 + 119)/22 489 × 100] is 99.24%. Calculating the same for six AT-AC pairs we extrapolate 12 cases and the general frequency should be 0.05%. Finally, the two non-canonical splice pairs can be extrapolated to four cases and will account for the rest of 0.02%. Although these estimates are rather crude, they are more likely reflecting the real situation than the numbers of non-canonical splice sites received from GenBank annotated data.

For GT-AG and for GC-AG populated splice pairs we now build donor and acceptor splice site weight matrices that can be used in gene prediction programs (Table 3). The donor and acceptor matrices for GT-AG pairs are very similar to the matrices constructed earlier on much smaller data sets (32). This is consistent with the growing number of known splice sites.

The GC-AG splice group has several interesting features. The first is its relatively high frequency (0.56% of all EST-supported splicing pairs belong to this type and 0.69% is the final estimated frequency of this group). It means that, on average in any 200 donor splice sites, one should be GC. For the first time we have obtained the weight (or frequency) matrix for this type of splice site (Table 3). It can be seen that this matrix shows a significantly higher degree of conservation in relation to the canonical donor matrix. This observation is in agreement with earlier investigations of GC donor sites consensus sequence (5). It provides the possibility to implement this information in gene prediction programs without generating many false positive predictions. Note that the characteristics of acceptor

AT-AC			HTG information
AC002397	**TGCCAAGATG**|atatccttgtgt	ctgtctgctcac|**CTTGGAGAAG**	Rodent
AC004976	**GAAAGAACCC**|atatcctttctg	actacttcatac|**AAAACAGTCA**	AT-AC
AF136179	**TATGGTAGAG**|atatcctttact	actgtttcggac|**ATTGACCAAA**	Rodent
AL021578	**ACGCTGAACC**|atatcctttggg	ttaaccgctcac|**TGGCCCAGCT**	AT-AC
L10295	**ATTGGTGAAG**|atatccttttag	aatcattactac|**ATGTGAATCC**	Rodent
U39892	**AGATTAGAGA**|atatcctttctt	aactgccagcac|**ATTTTGTCAG**	Rodent
U47924	**TGCCAAGATG**|atatccttctgc	aaccctcctcac|**CTTGGAGAAG**	AT-AC
U53004	**GGAAGTGGTC**|atatccttcctg	aactctgcacac|**GAAGCTCACG**	AT-AC

GT-AC			HTG information
AF102139	**TGATGACATG**|gtgcctgggaac	tcctcctcttac|**AACCTGCTGG**	GT-AG
U18671	**GGAGCTGAAA**|gtgagtgaaaat	ttctccccacac|**ACGGACACCC**	GT-AG
X14615	**CACTGCACAG**|gtgcccagacac	tcctgcttccac|**GGTCCTGGGC**	Pseudogene
X59520	**GTCCGCAAAG**|gtaagggtgtcc	ttctccctgcac|**CTCCTTCTGG**	Rodent

GT-CG			HTG information
M12381	**AGCAAGGGCG**|gtgagtgacccc	ggggtcccgacg|**GCTCTCACAC**	Rodent
M25133	**ATGCTGACAG**|gtgagcctggtg	ttttttcttccg|**ATCGGGAGAA**	GT-AG
U57691	**CAGCTTCTCA**|gtaagttcaccc	catcaattgtcg|**GTGAATTTGA**	Rodent
U92868	**CCATCGCCAC**|gtgagt------	------ttgccg|**CTTTCAGATT**	GT-AG

GT-TG			HTG information
AJ005919	**TGTGTCAAAG**|gtgagaaacatc	tgcgctctgctg|**AAGGACTCTC**	Rodent
D12905	**TTCTTTAAGC**|gtgagtctcagg	tctccaccactg|**CACACTGGGA**	Rodent
D14553	**CATGGTGCAG**|gtgcgacggggc	tatgttcccttg|**TCACCCATGT**	Rodent
U28932	**CTGGCACCAG**|gtgagtccagac	tctcttcccctg|**CTGGAGAACA**	Rodent
U70992	**AGCGCAGAGG**|gtgagtgcagac	ttttctttttttg|**CTATTATTCA**	GT-AG
X61600	**AGCTTATGAG**|gtacaatttgtg	tttgtattgatg|**GATTGAGGAG**	Rodent
X91922	**TACCAGACAA**|gtaggtgccttc	ctagaactagtg|**GATCCCCCTC**	GT-AG
Z93241	**GGTAGCGGAG**|gtggtgtttgtg	gaggagaaggtg|**GTGGGCTAGT**	GenBank HTG

AT-AG			HTG information
AB011399	**TGAGGAGCAG**|attgattgataa	ttacaggagcag|**ACGCCTCCGC**	GenBank HTG
AC006942	**CCAGACTCAT**|atcctctcaggc	tggcgggcacag|**CTGGCTGTGC**	AT-AC
AJ000412	**CACAACCTGT**|atgagtggggct	tgtgttctgcag|**GTGGAGAGAC**	Rodent
D87002	**CACTGCACTG**|atatgtgtcacc	cccatgccccag|**GCCATCATCT**	Pseudogene
L76571	**TTCAACCCCG**|ataaagaaactg	ctctgcccacag|**ATGTGCCAGG**	GT-AG
M13300	**CCCACTCGCT**|atcatcatgct-	--tctcccttag|**ATCATCATGC**	No intron
S71127	**ACCAGAAAGT**|ataagtgtctga	cttcccctgcag|**CTCCAAGACC**	No data

GA-AG			HTG information
AJ010070	**TACTATGGAG**|gacagtggttct	gtgatacaccag|**ATATTTCGTC**	GT-AG
D42128	**GTTTGGCAGG**|gaaggcagggta	tttctgaggtag|**GACCTTGTGA**	Rodent
L33779	**ACTTGGTGAT**|gaaaccattaga	aaacaataaaag|**AAGGTACAGT**	Rodent
M25133	**CTTGCTGAAG**|gacatgggggcg	cctgatgctcag|**GACACCCAGA**	GT-AG
M88067	**CAGGACTCAG**|gatgaggggggg	tctggtctgcag|**GTACTGCGGT**	Rodent
U28054	**GACCCCCCAG**|gataggagttgg	cccatcctacag|**ACCAGGTGCA**	GT-AG
U57999	**TCTGGCCACC**|gagctcgggcaa	gttttcctttag|**GCTCTGACCA**	Rodent
U77068	**GTGGAATGGA**|gagagaagctct	ttcttttctcag|**GAGCACAGTC**	Mammal

Figure 6. Small annotated and EST-supported non-canonical splice pair groups (without shifted dinucleotides).

splice site are very similar to the canonical pairs. The Fgenesh HMM-based gene prediction program (38) has been modified to predict genes with canonical as well as with GC-donor splice sites. This version of Fgenesh can identify non-standard GC-exons with approximately the same level of false positive predictions as the original program (http://genomic.sanger.ac.uk/gf/gf.shtml).

For the AT-AC pair we have constructed donor and acceptor site consensus sequences very similar to ones described before (20). In the acceptor sites, a slight bias to C or T upstream of the junction point (resembling the poly-C/T tract typical for canonical splice sites) was observed. In addition, all AT-AC human EST and HTG supported cases had consensus specific to AT-AC pairs described above, but we cannot identify any case of AT-AC with GT-AG consensus properties, noticed in the early works (24). However, the number of verified examples here is very small to draw the conclusion.

Despite the observations that most annotated non-canonical splice sites (excluding the above-mentioned three groups) are likely to have emerged from sequencing errors, we present sequences near each case of six non-canonical splice pairs, which do not contain shifted standard conserved dinucleotides (Fig. 6). A close look at these examples shows that the sequence composition around AT-AG, GA-AG, GT-AC, GT-CG and GT-TG is very similar to that observed for canonical pairs, but with a significantly lower level of conservation.

DISCUSSION

Genomic information is growing faster every day, but unfortunately the proportion of experimentally confirmed data is decreasing, which in turn raises the complexity of extraction of useful information. InfoGene database (28) provides us with one of the most complete gene centered genomic databases, with practically all coherent information that can be obtained

from the GenBank feature tables. We can retrieve the necessary information without looking at many GenBank entries, where the information about a particular gene might be stored. From InfoGene we have obtained practically all splice site pairs that can be extracted automatically for all mammals. The important step in subsequent analysis is the verification of these pairs using ESTs and correction of possible annotation errors.

Splice junctions verification by EST sequences (30) is a good procedure to check the accuracy of annotation, but as was observed in our work, is insufficient. A disadvantage of this approach is the absence of additional information about intron sequences as well as the high error rate in EST sequences. This motivated us to add another method to verify annotations based on comparisons between GenBank annotated genes and HTGs. In many cases HTGs provide independent sources of information. This approach could be applied as long as genome projects produce more new sequences.

GC-AG pairs were noticed as splice pair variant relatively early, but their frequency in respect to canonical pairs was not known and their characteristics were only studied for a very small set of mixed cases [26 cases, from soybean to human, (5)]. Here we characterized the properties of these pairs for mammals using a clean data set of 126 examples. GC-AG pairs could be taken into account in gene prediction programs, because while the proportion of total GC dinucleotides in respect to GC in splice sites is bigger than for GT dinucleotides, the conservation level of this donor site is higher. Therefore, gene finding approaches using standard GT-AG splice sites can potentially predict accurately ~97% of genes (assuming four exons per gene). Including GC-AG introns can increase this value to 99%.

After HTG verification practically all human non-canonical splice sites were reclassified as GT-AG, GC-AG and AT-AC, except two cases (GT-GG and TT-AG). In principle, all of these sites are supported by available sequences. However, the reference paper for U01337 is old, includes only cDNA characterization and contains no information about the genomic sequence. There is no reference at all to non-canonical pair from AF109620.

There is a possibility that HTG-supported non-canonical splice junctions actually belong to the annotated pseudogenes and the functional genes were not sequenced yet. Also, a very similar gene can produce an EST cross-supporting this junction or the error may appear due to incorrect EST sequencing (see the AF109620 example). There is only one EST supporting the AF109620 junction (Fig. 5c), and by deleting a 'g' in the donor or acceptor junction we can recover a GT-AG pair. It will be very interesting to investigate experimentally if there are annotation errors in these splice pairs. If these cases resolve, GT-AG, GC-AG and AT-AC remain as the only possible splice pairs.

Shapiro and Senapathy (6) reported one GG-AG and three CT-AG pairs in homologous genes. Applying our classification, we found the GG-AG case among EST-supported non-canonical pairs, but it could not be verified using human HTG because it is a mouse gene. For three CT-AG pairs we also recovered the mouse gene and they were unsupported by EST sequences.

Jackson (5) reported two GT-CG pairs for *Drosophila* genes and two GT-TG pairs for the same gene (*GS alpha*) of *Drosophila* and human. He indicates that all these genes are involved in alternative splicing. We found that the human *GS alpha* gene was classified in our set as ambiguous and removed from the investigated group. A closer examination of this case shows that a canonical splice site is clearly supported by ESTs and 3 nucleotides upstream was the annotated non-canonical splice site, also clearly supported by ESTs. This case was of particular interest because it falls in our discarded set of canonical and non-canonical splice sites, with ambiguous positions of splice junctions. Examining this group we found some examples of putative alternative positions of splice junctions at a very close distance (data not shown). At least one of these sites was canonical and both were supported by ESTs. It has been observed that mutations in position +5 of some donor sites produce less efficiency in splice cut specification and generate some other close cut points in addition to normal ones (39,40). Analysis of sequence characteristics around this alternative cut shows that they have no properties of cryptic splice sites, indicating that these alternative cut points are probably used as a consequence of such mutation and not as a consequence of the cryptic splice sites. Therefore, it can be postulated that in some circumstances the splicing process uses a canonical pair for the site recognition and in addition to the normal junction it can cut in close 'parasitic' sites, without splicing properties *per se*. It is interesting to note that these splice junctions tend to be in the same reading frame (data not shown). It seems that the additional cut position is tolerated if at least the coding frame is maintained. Although this kind of alternative splicing will generate very similar proteins, it could be a mechanism to introduce some variability into the system.

We have obtained eight classes of EST-supported splice pairs. Applying HTG on a small subset of these sites, only the GT-AG, GC-AG and AT-AC pairs were verified, except two cases, which could also be some kind of unspecified errors.

In some examples in Jackson's paper (5) and in our own data, we often observe alternative splicing variants, where some unusual non-canonical splice sites are located very close to a canonical one. It is possible that the only GT-AG, GC-AG and AT-AC pairs can recruit the splicing machinery effectively and the other non-canonical pairs could function exclusively in association with a canonical pair, which shares its properties with the neighbour, as some kind of parasitic splice sites. To verify this hypothesis we should additionally investigate EST-supported alternative splicing sites observed at a close distance.

ACKNOWLEDGEMENTS

We thank R. Guigó for very helpful comments. This work was supported by the Wellcome Trust grant to V.S. and grant FPI95 from Ministerio de Educación y Ciencia to M.B.

REFERENCES

1. Breathnach,R., Benoist,C., O'Hare,K., Gannon,F. and Chambon,P. (1978) *Proc. Natl Acad. Sci. USA*, **75**, 4853–4857.
2. Breathnach,R. and Chambon,P. (1981) *Annu. Rev. Biochem.*, **50**, 349–393.
3. Mount,S. (1981) *Nucleic Acids Res.*, **10**, 459–472.
4. Hodge,M.R. and Cumsky,M.G. (1989) *Mol. Cell. Biol.*, **9**, 2765–2770.
5. Jackson,I.J. (1991) *Nucleic Acids Res.*, **19**, 3795–3798.
6. Shapiro,M.B. and Senapathy,P. (1987) *Nucleic Acids Res.*, **15**, 7155–7174.
7. Wieringa,B., Meyer,F., Reiser,J. and Weissmann,C. (1983) *Nature*, **301**, 38–43.
8. Xue,J. and Rask,L. (1995) *Plant Mol. Biol.*, **29**, 167–171.
9. Quan,F. and Forte,M.A. (1990) *Mol. Cell. Biol.*, **10**, 910–917.
10. Feng,G.H., Bailin,T., Oh,J. and Spritz,R.A. (1997) *Hum. Mol. Genet.*, **6**, 793–797.

11. Hall,S.L. and Padgett,R.A. (1994) *J. Mol. Biol.*, **239**, 357–365.
12. Hall,S.L. and Padgett,R.A. (1996) *Science*, **271**, 1716–1718.
13. Kohrman,D.C., Harris,J.B. and Meisler,M.H. (1996) *J. Biol. Chem.*, **271**, 17576–17581.
14. Kolossova,I. and Padgett,R.A. (1997) *RNA*, **3**, 227–233.
15. Wu,Q. and Krainer,A.R. (1996) *Science*., **274**, 1005–1008.
16. Wu,Q. and Krainer,A.R. (1997) *RNA*, **3**, 586–601.
17. Wu,Q. and Krainer,A.R. (1998) *RNA*, **4**, 1664–1673.
18. Yu,Y.T. and Steitz,J.A. (1997) *Proc. Natl Acad. Sci. USA*, **94**, 6030–6035.
19. Nilsen,T.W. (1994) *Cell*, **78**, 1–4.
20. Sharp,P.A. and Burge,C.B. (1997) *Cell*, **91**, 875–879.
21. Tarn,W.Y. and Steitz,J.A. (1996) *Cell*, **84**, 801–811.
22. Tarn,W.Y. and Steitz,J.A. (1996) *Science*, **273**, 1824–1832.
23. Tarn,W.Y. and Steitz,J.A. (1997) *Trends Biochem. Sci.*, **22**, 132–137.
24. Burge,C.B., Padgett,R.A. and Sharp,P.A., (1998) *Mol. Cell*, **2**, 773–785.
25. Benson,D.A., Boguski,M.S., Lipman,D.J., Ostell,J., Ouellette,B.F., Rapp,B.A. and Wheeler,D.L. (1999) *Nucleic Acids Res.*, **27**, 12–17.
26. Stoesser,G., Tuli,M., Lopez,R. and Sterk,P. (1999) *Nucleic Acids Res.*, **27**, 18–24.
27. Penotti,F.E. (1991) *J. Theor. Biol.*, **150**, 385–420.
28. Solovyev,V.V. and Salamov,A.A. (1999) *Nucleic Acids Res.*, **27**, 248–250.
29. Boguski,M.S., Lowe,T.M. and Tolstoshev,C.M. (1993) *Nature Genet.*, **4**, 332–333.
30. Thanaraj,T.A. (1999) *Nucleic Acids Res.*, **27**, 2627–2637.
31. Altschul,S.F., Madden,T.L., Schaffer,A.A., Zhang,J., Zhang,Z., Miller,W. and Lipman,D.J. (1997) *Nucleic Acids Res.*, **25**, 3389–3402.
32. Senapathy,P., Sahpiro,M. and Harris,N. (1990) *Methods Enzymol.*, **183**, 252–278.
33. Dietrich,R., Incorvaia,R. and Padgett,R. (1997) *Mol. Cell*, **1**, 151–160.
34. Wu,Q. and Krainer,A.R. (1999) *Mol. Cell. Biol.*, **19**, 3225–3236.
35. Gunning,P., Ponte,P., Blau,H. and Kedes,L. (1983) *Mol. Cell. Biol.*, **3**, 1985–1995.
36. Hamada,H., Petrino,M.G. and Kakunaga T. (1982) *Proc. Natl Acad. Sci. USA*, **79**, 5901–5905.
37. Beck,T.W., Huleihel,M., Gunnell,M., Bonner,T.I. and Rapp,U.R. (1987) *Nucleic Acids Res.*, **15**, 595–609.
38. Salamov,A. and Solovyev,V. (2000) *Genome Res.*, **10**, 516–522.
39. Fouser,L.A. and Friesen,J.D. (1986) *Cell*, **45**, 81–93.
40. Parker,R. and Guthrie,C. (1985) *Cell*, **41**, 107–118.