D188–D192 *Nucleic Acids Research, 2007, Vol. 35, Database issue* *Published online 16 November 2006*
doi:10.1093/nar/gkl762

TassDB: a database of alternative tandem splice sites

Michael Hiller*, Swetlana Nikolajewa[1], Klaus Huse[2], Karol Szafranski[2], Philip Rosenstiel[3], Stefan Schuster[1], Rolf Backofen and Matthias Platzer[2]

Institute of Computer Science, Chair for Bioinformatics, Albert-Ludwigs-University Freiburg, Georges-Koehler-Allee 106, 79110 Freiburg, Germany, [1]Department of Bioinformatics, Friedrich-Schiller-University Jena, Ernst-Abbe-Platz 2, 07743 Jena, Germany, [2]Genome Analysis, Leibniz Institute for Age Research—Fritz Lipmann Institute, Beutenbergstrasse 11, 07745 Jena, Germany and [3]Institute of Clinical Molecular Biology, Christian-Albrechts-University Kiel, Schittenhelmstrasse, 12, 24105 Kiel, Germany

Received July 14, 2006; Accepted September 26, 2006

ABSTRACT

Subtle alternative splice events at tandem splice sites are frequent in eukaryotes and substantially increase the complexity of transcriptomes and proteomes. We have developed a relational database, TassDB (TAndem Splice Site DataBase), which stores extensive data about alternative splice events at GYNGYN donors and NAGNAG acceptors. These splice events are of subtle nature since they mostly result in the insertion/deletion of a single amino acid or the substitution of one amino acid by two others. Currently, TassDB contains 114 554 tandem splice sites of eight species, 5209 of which have EST/mRNA evidence for alternative splicing. In addition, human SNPs that affect NAGNAG acceptors are annotated. The database provides a user-friendly interface to search for specific genes or for genes containing tandem splice sites with specific features as well as the possibility to download large datasets. This database should facilitate further experimental studies and large-scale bioinformatics analyses of tandem splice sites. The database is available at http://helios.informatik.uni-freiburg.de/TassDB/.

INTRODUCTION

Alternative splicing is a very important step during pre-mRNA processing. As most of the human genes with multiple exons express more than one transcript, alternative splicing is considered to be a major mechanism for producing a complex proteome from a limited number of genes (1). The different transcripts of one gene can be translated into functionally different protein isoforms (2) or can be degraded by nonsense-mediated mRNA decay (3). The regulation of alternative splicing plays a role in several important processes such as the formation and function of synapses (4), axon guidance in Drosophila (5,6) and T-cell activation (7). Furthermore, defects in alternative splicing are causative for a number of human diseases (8,9) and thought to contribute to cancer development (10). Thus alternative splicing is also of therapeutic interest (11).

While much research focused on larger alternative splice events such as exon skipping, it recently became clear that numerous alternative splice events result in only subtle changes of the mRNA and of the protein (12–14). The most widespread type is the alternative splicing at acceptor sites with the pattern NAGNAG (N stands for A, C, G, or T, throughout the paper we write T instead of U also when referring to an RNA sequence) (12,15,16). In such a motif, both AGs represent potential alternative acceptor sites which result in transcripts that differ by only 3 nt (the NAG). About 6% of all human acceptors are NAGNAG acceptors. Based on expressed sequence tag (EST)/mRNA data 16% of all NAGNAGs and noteworthy 39% of the tandem acceptors with a HAGHAG pattern (also denoted 'plausible' NAGNAGs, H stands for A, C, or T) are currently known to be alternatively spliced. Furthermore, we recently found evidence for alternative splicing at donor splice sites with the motifs GTNGTN, GCNGTN and GTNGCN (denoted as GYNGYN donors, Y stands for C or T) where both GT/GC donors are used (17). We denote a tandem splice site as confirmed if the usage of both splice sites is represented by at least one EST/mRNA and unconfirmed otherwise. Although the term 'tandem splice site' refers to any pair of neighboring splice sites, in our database we collected data about NAGNAG acceptors and GYNGYN donors.

Apart from their frequency, subtle alternative splice events are of interest since several cases are known to result in

*To whom correspondence should be addressed. Tel: +49 761 203 8254; Fax: +49 761 203 7462; Email: hiller@informatik.uni-freiburg.de

functionally different protein isoforms (16,18–22) and alternative NAGNAG splicing in the untranslated region (UTR) can affect the translational efficiency (23). Moreover, the effect for the protein might be drastic since a premature stop codon can be created (12,17). Many NAGNAG acceptors are conserved between human and mouse and the ratio of the two splice forms can be highly controlled in a tissue-specific manner (12,16,24). Furthermore, SNPs that affect a NAGNAG acceptor can be relevant for human disease as demonstrated for the *ABCA4* gene (25) and suggested for many other genes (26).

While previous databases on alternative splicing do not store such subtle splice events (27–29), recent databases contain confirmed tandem splice sites (30–32). However, they do not contain unconfirmed tandem splice sites and do not allow to search for tandem splice sites with specific features.

To facilitate further experimental studies as well as large-scale bioinformatics analyses of tandem splice sites, we have developed a relational database, TassDB (TAndem Splice Site DataBase), which provides large collections of GYNGYN donors and NAGNAG acceptors in eight species. Since these subtle splice events can easily be overlooked in experimental systems (a 3 nt difference between two bands is barely visible on an agarose gel) and additional alternative splice events are likely to be missed in current EST data, TassDB also stores unconfirmed tandem splice sites. A user interface allows to search for genes of interest and to get all relevant information about the respective tandem splice sites. It is also possible to search for genes harboring GYNGYNs/NAGNAGs with specific features. Finally, TassDB annotates NAGNAGs that are affected by a SNP and also contains 51 tandem acceptors whose NAGNAG pattern can only be observed in another SNP-allele and not in the human genome reference sequence.

TassDB

Data

As alternatively spliced NAGNAG acceptors and GYNGYN donors occur in a large number of lineages including vertebrates, flies and nematodes, TassDB stores information about the tandem splice sites of eight species: *Homo sapiens*, *Canis familiaris*, *Mus musculus*, *Rattus norvegicus*, *Gallus gallus*, *Danio rerio*, *Drosophila melanogaster* and *Caenorhabditis elegans*. Our annotation pipeline is based on transcript-to-genome mappings taken from the UCSC genome browser (33). Apart from the RefSeq annotation that was used for all species, we additionally used Ensembl transcripts for human, rat, chicken, zebrafish, the UCSC 'knownGene' set for human, mouse, rat (34), flyBase transcripts for *Drosophila* and wormBase transcripts for *C.elegans*. The exon–intron structure as well as the annotation of the open reading frame was taken from the UCSC annotation. To identify alternatively spliced NAGNAGs and GYNGYNs, we used BLAST against all ESTs and mRNAs from the respective species as described in Refs (12,17). All transcripts and expressed sequences were downloaded in April 2006. The SNPs that affect NAGNAG acceptors were taken from (26).

Database design

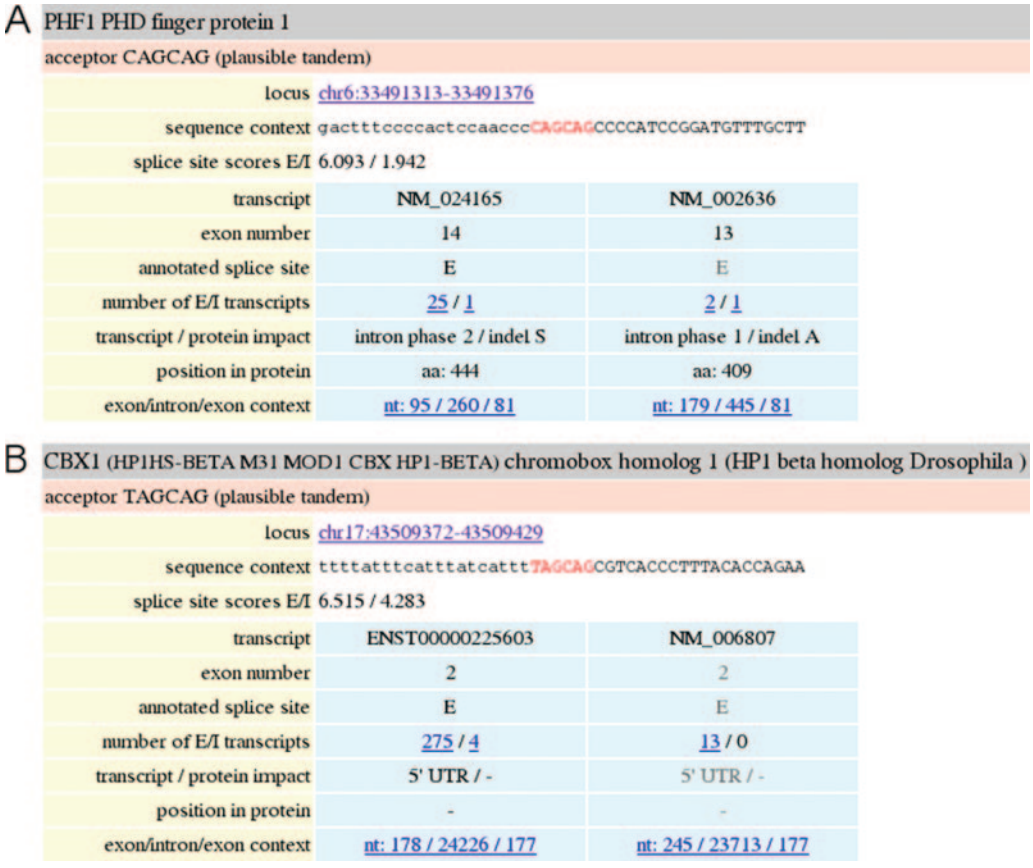
The primary aim of TassDB is to provide information that is specific for the tandem splice site and the putative alternative splice event. Thus, we collected the following data: the splice site motif, its genomic locus, its location in the transcript (5′/3′-UTR or CDS with intron phase 0/1/2), the impact of the splice event on the protein, the sequences and length of the up-/downstream exon and the intron, and information about the ESTs/mRNAs that indicate usage of one of the two splice sites. As the degree of similarity of a splice site to the overall consensus is an important criterion to distinguish alternatively from non-alternatively spliced NAGNAG acceptors (15), we also computed the maximum entropy scores for both splice sites in a tandem (35).

The basic design of TassDB was driven by the idea to separate splice site specific data from transcript specific data. For example, the GYNGYN/NAGNAG motif, the genomic locus and the splice site scores are independent of transcript annotation. However, features such as intron phase, protein impact and EST confirmation depend on the annotation and the exon–intron structure of the transcript. Thus, one tandem splice site can have multiple transcript specific data. For example, the intron 13 of the *PHF1* gene that contains a CAGCAG acceptor is in intron phase 2 according to the annotation of NM_024165. Due to skipping of the upstream 95 nt exon, this intron is in phase 1 according to the annotation of another transcript NM_002636. Thus, the protein impact of the CAGCAG is the insertion/deletion (indel) of a Ser in NM_024165 but indel Ala in NM_002636 (Figure 1A). Another example is the CAGCAG acceptor of intron 1 of the *CBX1* gene having two alternative first exons (represented by the transcripts ENST00000225603 and NM_006807). We found EST evidences for alternative CAGCAG splicing if the upstream first exon is used (ENST00000225603) but not if the downstream first exon is used (NM_006807) (Figure 1B). Whether this is a biological phenomenon or simply the consequence of a lack of sufficient EST data, deserves further research.

User interface

The most frequent use of TassDB might be a search for tandem splice sites of a given gene. To this end, TassDB provides a quick search interface where a user only specifies a gene symbol or a transcript accession and gets the entire information of both confirmed and unconfirmed GYNGYNs and NAGNAGs for this gene.

A more advanced task is to select tandem splice sites with specific features for further experimental or computational analysis. To this end, TassDB provides an advanced search interface where the user can restrict the search to GYNGYNs or NAGNAGs with the following features: (i) pattern of the splice site, (ii) number of ESTs/mRNAs that match both splice forms, (iii) location in the UTR or in the coding sequence (CDS) (as well as specific intron phases), and (iv) protein impact (Figure 2). Thus, it is easy to formulate queries such as: (i) Show all confirmed NAGNAGs that result in single amino acid events. (ii) Show all tandem splice sites where both splice forms are represented by at least three ESTs/mRNAs and that are located in the



A PHF1 PHD finger protein 1

acceptor CAGCAG (plausible tandem)

locus	chr6:33491313-33491376	
sequence context	gactttccccactccaacccCAGCAGCCCCATCCGGATGTTTGCTT	
splice site scores E/I	6.093 / 1.942	
transcript	NM_024165	NM_002636
exon number	14	13
annotated splice site	E	E
number of E/I transcripts	25 / 1	2 / 1
transcript / protein impact	intron phase 2 / indel S	intron phase 1 / indel A
position in protein	aa: 444	aa: 409
exon/intron/exon context	nt: 95 / 260 / 81	nt: 179 / 445 / 81

B CBX1 (HP1HS-BETA M31 MOD1 CBX HP1-BETA) chromobox homolog 1 (HP1 beta homolog Drosophila)

acceptor TAGCAG (plausible tandem)

locus	chr17:43509372-43509429	
sequence context	ttttatttcatttatcatttTAGCAGCGTCACCCTTTACACCAGAA	
splice site scores E/I	6.515 / 4.283	
transcript	ENST00000225603	NM_006807
exon number	2	2
annotated splice site	E	E
number of E/I transcripts	275 / 4	13 / 0
transcript / protein impact	5' UTR / -	5' UTR / -
position in protein	-	-
exon/intron/exon context	nt: 178 / 24226 / 177	nt: 245 / 23713 / 177

Figure 1. (**A**) Example of a NAGNAG acceptor whose protein impact differs in the annotation of two transcripts. (**B**) Example of a NAGNAG acceptor that is only confirmed according the exon-intron structure of one but not a second transcript.

5′-UTR. (iii) Show all confirmed GYNGYNs where one donor has a GC dinucleotide. Additionally, the search can be restricted to certain genes. The result of the search consists of two parts: (i) summary table of affected genes and their number of tandem splice sites and (ii) detailed tables with information about the tandem splice sites. The detailed result tables also provide links to the ESTs/mRNAs for both splice forms as well as links to the UCSC genome browser. If the transcript specific data differ between transcripts, TassDB will show detailed result tables with more than two columns (Figure 1). Features that differ between transcripts are shown in black while those that are equal in all transcripts are shown in grey colour.

As SNPs that affect NAGNAG acceptors are predictive for variation in alternative splicing, TassDB also annotates the SNP data found in our previous study (26) including 51 polymorphic tandem acceptors whose NAGNAG pattern is not visible in the genome reference sequence. Such an example is the *APPBP1* gene where acceptor pattern of intron 6 is AAACAG in the human reference genome sequence. The SNP rs363209 leads to a second allele with an AAGCAG pattern, thus this G-allele creates a novel tandem acceptor. TassDB always contains the sequence with the NAGNAG acceptor (in this case the G-allele). Links to dbSNP are provided.

Finally, TassDB provides an interface where the user can send an arbitrary SQL select query to the database. The relational database schema with table and column names are given on the web page. Thus, TassDB can be used to retrieve large datasets for further computational analysis of tandem splice sites.

Statistics

The current statistics of TassDB are shown in Table 1. It is evident that tandem splice sites are widespread in all eight species and that with the exception of *C.elegans* for NAGNAGs numerous of them are already known to be alternatively spliced.

FUTURE DIRECTIONS

As tandem splice sites seem to occur in all species allowing alternative splicing, it would be desirable to include information for other species. In the future, we plan to include *Bos taurus*, *Xenopus tropicalis*, *Takifugu rubripes* and *Anopheles gambiae* as more ESTs and annotated genes become available. Furthermore, information about conservation of these splice events in other species as well as more data about the protein impact may be included. Since the database was designed to store information about any tandem splice site, an extension to tandem splice sites that are more than 3 nt apart is conceivable.

Figure 2. Screenshot of the advanced search interface: one can search for genes harboring tandem splice sites with a specific pattern [such as GTNGCN for tandem donors or HAGHAG (plausible) for tandem acceptors], a certain number of ESTs/mRNAs matching both splice forms, a location in the UTR or CDS and a specific protein impact. Furthermore, one can search only for GYNGYNs or NAGNAGs.

Table 1. Summary of the current content of TassDB

	No. of transcripts[a]	No. of ESTs/mRNAs[b]	GYNGYN Total	Confirmed		NAGNAG Total	Confirmed	
Homo sapiens	73 923	7 948 538	10 995	141	1.28%[c]	11 964	1945	16.3%[c]
Canis familiaris	674	361 114	312	0	—	268	15	5.6%
Mus musculus	35 897	4 952 715	8723	91	1.04%	9815	1487	15.2%
Rattus norvegicus	38 347	900 354	7916	20	0.25%	8658	415	4.8%
Gallus gallus	27 996	618 736	8538	19	0.22%	9323	377	4.0%
Danio rerio	41 956	879 246	9334	24	0.26%	11 234	368	3.3%
Drosophila melanogaster	39 733	422 535	3752	33	0.88%	1842	208	11.3%
Caenorhabditis elegans	45 410	327 081	7054	32	0.45%	4826	34	0.7%
Sum	303 936	16 410 319	56 624	360	0.64%	57 930	4849	8.4%

[a]Total number of transcripts used for detecting tandem splice sites.
[b]Total number of ESTs and mRNAs.
[c]Number of confirmed tandems/number of all tandems.

ACKNOWLEDGEMENTS

We thank Stefan Jankowski for technical help and Anke Busch for critical reading of the manuscript. This work was supported by grants from the German Ministry of Education and Research to P.R. (01GS0426) and to M.P. (01GR0504, 0313652D) as well as from the Deutsche Forschungsgemeinschaft (SFB604-02) to M.P. Funding to pay the Open Access publication charges for this article was provided by the Albert-Ludwigs-University Freiburg.

Conflict of interest statement. None declared.

REFERENCES

1. Maniatis,T. and Tasic,B. (2002) Alternative pre-mRNA splicing and proteome expansion in metazoans. *Nature*, **418**, 236–243.
2. Stamm,S., Ben-Ari,S., Rafalska,I., Tang,Y., Zhang,Z., Toiber,D., Thanaraj,T.A. and Soreq,H. (2005) Function of alternative splicing. *Gene*, **344**, 1–20.
3. Lewis,B.P., Green,R.E. and Brenner,S.E. (2003) Evidence for the widespread coupling of alternative splicing and nonsense-mediated mRNA decay in humans. *Proc. Natl Acad. Sci. USA*, **100**, 189–192.
4. Ule,J., Ule,A., Spencer,J., Williams,A., Hu,J.S., Cline,M., Wang,H., Clark,T., Fraser,C., Ruggiu,M. *et al.* (2005) Nova regulates brain-specific splicing to shape the synapse. *Nature Genet.*, **37**, 844–852.
5. Neves,G., Zucker,J., Daly,M. and Chess,A. (2004) Stochastic yet biased expression of multiple Dscam splice variants by individual cells. *Nature Genet.*, **36**, 240–246.
6. Wojtowicz,W.M., Flanagan,J.J., Millard,S.S., Zipursky,S.L. and Clemens,J.C. (2004) Alternative splicing of *Drosophila* Dscam generates axon guidance receptors that exhibit isoform-specific homophilic binding. *Cell*, **118**, 619–633.
7. Lynch,K.W. (2004) Consequences of regulated pre-mRNA splicing in the immune system. *Nature Rev. Immunol.*, **4**, 931–940.
8. Faustino,N.A. and Cooper,T.A. (2003) Pre-mRNA splicing and human disease. *Genes Dev.*, **17**, 419–437.
9. Garcia-Blanco,M.A., Baraniak,A.P. and Lasda,E.L. (2004) Alternative splicing in disease and therapy. *Nat. Biotechnol.*, **22**, 535–546.
10. Kalnina,Z., Zayakin,P., Silina,K. and Line,A. (2005) Alterations of pre-mRNA splicing in cancer. *Genes Chromosomes Cancer*, **42**, 342–357.
11. Sazani,P. and Kole,R. (2003) Therapeutic potential of antisense oligonucleotides as modulators of alternative splicing. *J. Clin. Invest.*, **112**, 481–486.
12. Hiller,M., Huse,K., Szafranski,K., Jahn,N., Hampe,J., Schreiber,S., Backofen,R. and Platzer,M. (2004) Widespread occurrence of alternative splicing at NAGNAG acceptors contributes to proteome plasticity. *Nature Genet.*, **36**, 1255–1257.
13. 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.
14. Zavolan,M., Kondo,S., Schonbach,C., Adachi,J., Hume,D.A., Hayashizaki,Y. and Gaasterland,T. (2003) Impact of alternative initiation, splicing, and termination on the diversity of the mRNA transcripts encoded by the mouse transcriptome. *Genome Res.*, **13**, 1290–1300.
15. Chern,T.M., van Nimwegen,E., Kai,C., Kawai,J., Carninci,P., Hayashizaki,Y. and Zavolan,M. (2006) A simple physical model predicts small exon length variations. *PLoS Genet.*, **2**, e45.
16. Tadokoro,K., Yamazaki-Inoue,M., Tachibana,M., Fujishiro,M., Nagao,K., Toyoda,M., Ozaki,M., Ono,M., Miki,N., Miyashita,T. *et al.* (2005) Frequent occurrence of protein isoforms with or without a single amino acid residue by subtle alternative splicing: the case of Gln in DRPLA affects subcellular localization of the products. *J. Hum. Genet.*, **50**, 382–394.
17. Hiller,M., Huse,K., Szafranski,K., Rosenstiel,P., Schreiber,S., Backofen,R. and Platzer,M. (2006) Phylogenetically widespread alternative splicing at unusual GYNGYN donors. *Genome Biol.*, **7**, R65.
18. Condorelli,G., Bueno,R. and Smith,R.J. (1994) Two alternatively spliced forms of the human insulin-like growth factor I receptor have distinct biological activities and internalization kinetics. *J. Biol. Chem.*, **269**, 8510–8516.
19. Lorkovic,Z.J., Lehner,R., Forstner,C. and Barta,A. (2005) Evolutionary conservation of minor U12-type spliceosome between plants and humans. *RNA*, **11**, 1095–1107.
20. Ray,D.W., Davis,J.R., White,A. and Clark,A.J. (1996) Glucocorticoid receptor structure and function in glucocorticoid-resistant small cell lung carcinoma cells. *Cancer Res.*, **56**, 3276–3280.
21. Rivers,C., Levy,A., Hancock,J., Lightman,S. and Norman,M. (1999) Insertion of an amino acid in the DNA-binding domain of the glucocorticoid receptor as a result of alternative splicing. *J. Clin. Endocrinol. Metab.*, **84**, 4283–4286.
22. Vogan,K.J., Underhill,D.A. and Gros,P. (1996) An alternative splicing event in the Pax-3 paired domain identifies the linker region as a key determinant of paired domain DNA-binding activity. *Mol. Cell. Biol.*, **16**, 6677–6686.
23. Joyce-Brady,M., Jean,J.C. and Hughey,R.P. (2001) gamma-glutamyltransferase and its isoform mediate an endoplasmic reticulum stress response. *J Biol Chem*, **276**, 9468–9477.
24. Akerman,M. and Mandel-Gutfreund,Y. (2006) Alternative splicing regulation at tandem 3′ splice sites. *Nucleic Acids Res.*, **34**, 23–31.
25. Maugeri,A., van Driel,M.A., van de Pol,D.J., Klevering,B.J., van Haren,F.J., Tijmes,N., Bergen,A.A., Rohrschneider,K., Blankenagel,A., Pinckers,A.J. *et al.* (1999) The 2588G→C mutation in the ABCR gene is a mild frequent founder mutation in the Western European population and allows the classification of ABCR mutations in patients with Stargardt disease. *Am. J. Hum. Genet.*, **64**, 1024–1035.
26. Hiller,M., Huse,K., Szafranski,K., Jahn,N., Hampe,J., Schreiber,S., Backofen,R. and Platzer,M. (2006) Single-nucleotide polymorphisms in NAGNAG acceptors are highly predictive for variations of alternative splicing. *Am. J. Hum. Genet.*, **78**, 291–302.
27. Lee,C., Atanelov,L., Modrek,B. and Xing,Y. (2003) ASAP: the alternative splicing annotation project. *Nucleic Acids Res.*, **31**, 101–105.
28. Pospisil,H., Herrmann,A., Bortfeldt,R.H. and Reich,J.G. (2004) EASED: Extended Alternatively Spliced EST Database. *Nucleic Acids Res.*, **32**, D70–D74.
29. Stamm,S., Riethoven,J.J., Le Texier,V., Gopalakrishnan,C., Kumanduri,V., Tang,Y., Barbosa-Morais,N.L. and Thanaraj,T.A. (2006) ASD: a bioinformatics resource on alternative splicing. *Nucleic Acids Res.*, **34**, D46–D55.
30. de la Grange,P., Dutertre,M., Martin,N. and Auboeuf,D. (2005) FAST DB: a website resource for the study of the expression regulation of human gene products. *Nucleic Acids Res.*, **33**, 4276–4284.
31. Holste,D., Huo,G., Tung,V. and Burge,C.B. (2006) HOLLYWOOD: a comparative relational database of alternative splicing. *Nucleic Acids Res.*, **34**, D56–62.
32. Nagasaki,H., Arita,M., Nishizawa,T., Suwa,M. and Gotoh,O. (2006) Automated classification of alternative splicing and transcriptional initiation and construction of visual database of classified patterns. *Bioinformatics*, **22**, 1211–1216.
33. Hinrichs,A.S., Karolchik,D., Baertsch,R., Barber,G.P., Bejerano,G., Clawson,H., Diekhans,M., Furey,T.S., Harte,R.A., Hsu,F. *et al.* (2006) The UCSC Genome Browser Database: update 2006. *Nucleic Acids Res.*, **34**, D590–D598.
34. Hsu,F., Kent,W.J., Clawson,H., Kuhn,R.M., Diekhans,M. and Haussler,D. (2006) The UCSC known genes. *Bioinformatics*, **22**, 1036–1046.
35. Yeo,G. and Burge,C.B. (2004) Maximum entropy modeling of short sequence motifs with applications to RNA splicing signals. *J. Comput. Biol.*, **11**, 377–394.